AN IMPROVED SOURCE MODEL FOR A LINEAR PREDICTION SPEECH SYNTHESIZER

BY

HWAI-TSU HU

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1993

To my parents,

and to my wife

ACKNOWLEDGMENTS

I am deeply grateful to my advisor and committee chairman, Dr. Donald G. Childers, who encouraged me to explore my ideas in the field of speech science. Throughout my investigation, Dr. Childers provided guidance, assistance and financial support. I would also like to thank Dr. L. W. Couch II, Dr. F. J. Taylor, Dr. J. C. Principe and Dr. M. C. K. Yang, for serving on my supervisory committee and advising me on various aspects of this dissertation. My colleagues at the Mind-Machine Interaction Research Center helped me in many ways. I appreciate their interest and contributions to my research. Finally, I would like to dedicate this dissertation to my wife, Yuh-Yeh, who patiently shared every struggle I experienced during the process, and to my parents, who never hesitated to deliver their love, encouragement and support.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

AN IMPROVED SOURCE MODEL FOR A LINEAR PREDICTION SPEECH SYNTHESIZER

By

Hwai-Tsu Hu

May, 1993

Chairman: Dr. D G. Childers
Major Department: Electrical Engineering

Though some progress has been made towards producing natural-sounding speech using the linear prediction (LP) techniques, an appropriate feature-based parametric excitation source has not been fully developed for this type of synthesizer. The intent of this research is to verify the importance of selected acoustic measures by means of LP analysis within the source-filter theory, and then use the deduced information to develop an LP synthesizer that is capable of synthesizing high-quality natural-sounding speech.

In order to carry out the requirements of this research, we divide the relevant issues into two separate but related phases. In the first phase we propose methods for isolating and extracting the acoustic features of vocal quality. Based upon a comprehensive speech production model, the LP analysis is used to estimate spectral properties of the speech signals and of the glottal source. Relevant source features, other than jitter and shimmer, are retrieved from the integrated residue. An algorithm is developed to extract the time domain characteristics of the vocal noise. Various aspects of such extracted noise are examined subsequently. To illustrate the above mentioned analysis techniques, the measured acoustic parameters of the proposed speech model for three voice types (modal, vocal fry, and

breathy) are provided as representative examples. It is anticipated that our findings will contribute to the understanding of the problems of modeling the excitation source and the LP synthesizer

In the second phase we propose a novel source model to simulate the residue signal in terms of the glottal phase characteristics Depending on the voicing condition of the analyzed speech, the excitation source is formulated as two separate codebooks, i.e., a glottal codebook for voiced segments and a stochastic codebook for unvoiced and silence segments. Methods for determining voicing intervals are presented, along with procedures for searching the codewords for the appropriate excitation. Since pitch synchronous schemes are preferred for speech synthesis, we describe procedures for identifying the instants of glottal closure and for interpolating the excitation pulses as well as the LP coefficients. Moreover, we account for the effects of vocal noise and source-tract interaction, which are generally ignored in most synthesizers Finally, a method for determining the voicing gain is given. This method also serves as an expository tool to explicate the relationship between the gain and power intensity Informal listening tests were used to evaluate the speech processing techniques. The listening tests revealed that the quality of synthetic speech was close to that of the original speech. The results indicate that our source model is able to characterize the glottal features and that the overall speech production model is quite adequate for high-quality synthesis.

CHAPTER 1
INTRODUCTION

Speech is a sophisticated skill that humans have developed for efficient communication. This skill transmits not only linguistic information but also acoustic features that convey the speaker's identity and other aspects of the speaker's physical and emotional state. Although our current knowledge is insufficient to unveil the linguistic rules that describe the phonatory system the mechanism of speech production is becoming comprehensible due to the advances in acoustic theory and computing technologies. Phonatory acoustics forms the basis for all present-day speech synthesizers. The increasing use of speech synthesizers in the marketplace has produced great demand for products that can generate "high-quality" speech. In fact, quality degradation with regard to existing speech synthesizers mostly results from unnatural-sounding characteristics, which are known to cause perceptual difficulties (Pisoni and Hunnicutt, 1980; Pisoni et al., 1983). The aim of this dissertation was to seek methods that could improve the naturalness of synthetic speech. We restricted our attention to the Linear Prediction (LP) technique because of its simplicity and accuracy in speech processing.

Following a brief introduction to speech production, we give an overview of some existing synthesis techniques. Then we provide the details of speech modeling and processing. This overview, associated with the basic knowledge of the phonatory system, facilitates our explanation of some efforts made to improve the naturalness of synthetic speech.

1.1 Speech Production Mechanism

Perhaps the easiest way to describe the speech production mechanism is to explain the physiological function of the anatomy of the human vocal system. In general, the speech

production system can be divided into two systems, namely, the excitation source and acoustic modulation, which shapes the excitation spectrum to form intelligible sounds (or phonemes)

1.1.1 Excitation Source

The lungs act as an air reservoir, expelling air up the trachea to the vocal folds During periods of voiced speech the vocal folds open and close in a quasi-periodic fashion, producing a pulsating airstream. During the periods of unvoiced speech, the vocal folds are held apart so that the airstream is less disturbed and can be considered as a steady turbulent source. The vocal folds have an important role in determining the characteristics of vocal quality. The oscillation of the vocal folds can be described by the aerodynamic myoelastic theory (Berg et al., 1957, Berg, 1958). Basically, the motion of the vocal folds is controlled by several interplaying forces that cause the abduction and adduction of the folds. When the subglottal pressure is built up to a certain level, the vocal folds are pushed apart and the air is then released through the glottis. The volume velocity of air passing through the glottis increases as the vocal folds keep opening. As the velocity increases beyond some threshold, pressure across the folds begins to drop and then results in a Bernoulli effect. This effect, in conjunction with the elastic resistance of the folds, initiates the adduction of the vocal folds at the time these two effects outweigh the subglottal pressure. When the vocal folds close, the subglottal pressure builds up again and the entire procedure repeats. Such a repetitive cycle is referred to as a pitch period, its reciprocal is denoted as the fundamental frequency

Noise generated by turbulence is another important source of speech production. The airflow emerging from the lungs can cause turbulent streaming while passing through a vocal aperture, which is either the vibrating vocal folds or a constriction along the vocal tract. Such turbulence ceases if the vocal aperture opens sufficiently or the airflow decreases. The possibility of turbulent flow is indicated by the value of the Reynolds number, which

characterizes the viscosity of the airstream as either laminar turbulent or somewhere in between (Acheson, 1990) With these kinds of characteristics, there is no doubt that the turbulence becomes an essential element in fricative, aspirative, plosive, whisper and breathy sounds. This fact necessitates the use of a noise source while synthesizing those particular sounds

1.1.2 Acoustic Modulation

The human vocal tract, extending from the glottis to the lips, can be considered as an acoustic tube of nonuniform shape varying as a function of time. Components that lead to this time varying change include the lips, jaw, tongue, velum and nasal cavity. During the periods of nonnasal sounds the velum closes off the nasal tract from the vocal tract. Thus, the acoustic tube only exhibits poles in its transfer function. When the velum is lowered, the vocal tract is acoustically coupled with the nasal tract, forming a pole-zero system As the tube varies the shape for different sounds, the resultant transfer function is such that it emphasizes certain frequency components of the glottal wave and/or de-emphasizes others The resonant peaks of the speech output due to the poles are referred to as formants, whereas the valleys due to the zeros are referred to as anti-formants.

1.2 Previous Research on Speech Production

The earliest efforts of speech research were directed to exploring the physiological nature of the human phonatory system. At that time, the speech synthesizers played a fundamental role in learning the process of speech production. The talking machine, designed by von Kempelen in 1791, contained a bellows which supplied air to a reed (Flanagan, 1972b), the bellows and the reed were obviously used to simulate the lungs and the vocal folds respectively A hand-varied resonator was provided to simulate the acoustic response of the vocal tract. This machine was reported to produce only a few vowels Modern speech synthesizers are electrical in nature. Technologies developed over this

century have come out with sophisticated techniques which greatly improved the quality of synthetic speech. Such a technological evolution accompanied with the emerging understanding of speech acoustics, gradually shifted the focuses and interests of speech synthesis to other applications. The most significant influence was the Vocoder invented by Dudley (Dudley, 1939), whose efforts spawned a subfield of communication engineering. Research in this subfield was aimed at the efficient encoding and transmission of speech information. The techniques of interest were directed toward obtaining acceptable quality at low bit rates, using reasonable computational resources in a real time environment. Research issues encompassed methods to improve quality, robustness, delay and complexity.

As speech synthesis techniques have continued to improve in recent years, many speech synthesizers have been employed to implement voice response systems for computers, which are called the "text-to-speech" techniques. Speech synthesis in the sense of "text-to-speech" means automatically producing voice response according to a text input. The capability of voice response offers possibilities for automatic information services, computer based instruction, talking aids for the vocally handicapped, and reading aids for the visually impaired.

1.3 Models for Speech Synthesis

This research is directed toward improving the speech production model. We define the term "speech analysis" as the procedure used to extract the speech production model parameters from the speech signal and "speech synthesis" as the procedure used to reproduce the acoustic speech signal by controlling and updating the appropriate parameters obtained from the speech analysis.

Modern speech synthesizers can be classified into two groups, one based on the Fourier transform methods and the other based on the linear source-filter model.

1.3.1 Fourier Model

The Fourier transform has traditionally been used to study speech signals because it provides a frequency domain analysis of the phonatory and auditory properties of speech signals. Using the Fourier model, the speech signal is analyzed using short-time Fourier analysis (STFA), while synthesis is carried out by an inverse transform (Allen, 1977, Allen and Rabiner, 1977). The term "short-time" implies that the speech spectrum is stationary over a short interval of time. This is a valid approach to speech processing because many psychoacoustic and physiological studies have shown that the human ear performs a type of short time spectral analysis of acoustic signals.

The channel vocoder is the oldest form of speech coding device that exploits Fourier analysis and synthesis (Dudley, 1939). This vocoder is constituted by several bandpass filters, each of which is employed to preserve the magnitude Fourier transform of the speech signal within a specific band. An additional channel is needed to transmit other information regarding the excitation, e.g., the voiced/unvoiced signal and the pitch period for voiced speech. Consequently, the concept of source excitation was incorporated into the configuration of the channel vocoder.

Another Fourier-based model that has experienced popularity is the phase vocoder (Flanagan and Golden, 1966). The major success of this technique originates from a polar representation of the Fourier transformation, i.e., phase and amplitude, which leads to an economy of transmission bandwidth. Unlike the channel vocoder, which neglects the phase spectrum, the phase vocoder exploits the phase information through the derivative of the phase spectrum. Furthermore, it provides flexibility for expanding and compressing the time scale through the manipulation of the instantaneous frequency. Emerging from a similar idea, a new class of models called "sinusoidal coders" were developed and have proliferated since the early 1980s (Hedelin, 1981; Almeida and Tribolet, 1982, Almeida and Silva, 1984, McAulay and Quatieri, 1984; Trancoso et al., 1990). For such coders, the speech signal

within each frame is represented by a superposition of sinusoids with time varying amplitude and frequencies:

$$s(t) = \sum_{k=1}^{n} a_k(t)\cos\phi_k(t), \qquad 0 < t < T \tag{1-1}$$

where n is the number of sinusoids, $a_k(t)$ is the amplitude of the kth sinusoids, $\phi_k(t)$ is the corresponding frequency and T is the frame length. The variation of amplitudes, a_k's, and phases, ϕ_k's, within a short interval is usually described by first- and third-order polynomials respectively as

$$a_k(t) = A_k + (t/T)B_k, \tag{1-2}$$

$$\phi_k(t) = c_{3k}t^3 + c_{2k}t^2 + c_{1k}t + c_{0k}. \tag{1-3}$$

These polynomials are then applied to an interpolation rule for the instantaneous values of amplitude and phase as well as frequency. With a 10 ms frame, speech quality obtained using this model is virtually indistinguishable from the original.

Among the sinusoidal coders, approaches for processing the Fourier-based parameters can be divided into two classes. Members in one class separate the pitch harmonics from the spectral envelope, and only apply the sinusoidal processing techniques to the harmonics. In other words, the a_k's in Eq. (1-1) are obtained by other means such as linear prediction and ceptrum analysis. Members in the other class, on the other hand, consider the a_k's as part of the results of Fourier analysis. Interestingly, the generation of noise excitation also exhibits two different forms, i.e., either white noise or a signal with random phases but constant amplitudes.

1.3.2 Source-filter Model

The source-filter model was developed by Fant in the late 1950s (Fant, 1959; Fant, 1960). In this model the speech signal is modeled as the filtered output of a network excited

by quasi-periodic pulses for voiced speech or by random noise for unvoiced speech. The transfer function of the network is defined as the ratio of the Laplace transform of the sound pressure from the lips of the speaker to the volume velocity of the airflow passing the vocal folds. In the sense of speech production, a speech signal inherits both characteristics of the source and the network. Formed based on the source-filter theory, speech synthesizers can be further classified into three categories, namely, the LP, formant and articulatory synthesizers.

1.3.2.1 Articulatory synthesizer

The articulatory synthesizer is a direct approach that simulates speech production and propagation from the viewpoint of anatomy and physiology. In order to describe the wave propagation by means of aerodynamic equations, we must specify such parameters as subglottal pressure, elasticity of the vocal folds and viscosity of the vocal tract, in addition to the movement of the articulator coordinates and the changes in the vocal fold configuration. As shown in Figure 1–1, although the overall computation can be broken down into a sequence of subsegments of constant cross-sectional areas, the complexity involved in this aerodynamic and mechanical system is still considerable. Therefore, researchers have attempted to convert the gross feature of vocal fold vibration into a model with acoustic parameters. Likewise, the vocal and nasal tracts were represented by an equivalent circuit such as an analog transmission line (Figure 1–2). Furthermore, the control system for driving the area function of the vocal tract was developed by matching the formant characteristics of the model to those of real speech.

1.3.2.2 Formant synthesizer

The development of the formant synthesizer is mainly based on the perceptual characteristics of the human auditory apparatus. In this type of synthesizer, the transfer

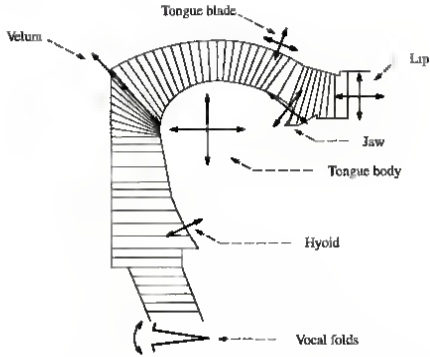


Figure 1–1 Articulatory model of human vocal tract and the associated control variables.

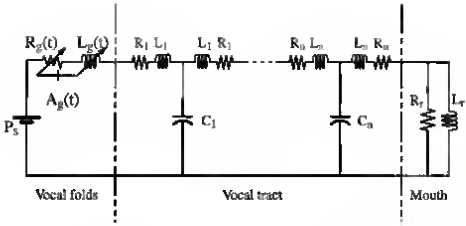


Figure 1–2. Equivalent circuit for the vocal system.

function is directly controlled by the use of resonant and anti-resonant filters whose center frequencies and bandwidths can be individually specified. The resonant filters can be connected either in parallel or in series so as to facilitate the production of both nasal and nonnasal sounds. An excitation model resembling the natural excitation is used to provide the source properties. Figure 1–3 shows the diagram of a typical formant synthesizer

1.3.2.3 LP synthesizer

The LP synthesizer consists of an excitation source and a time-varying all-pole filter (Figure 1–4). The all-pole filter determines the spectral envelope of the synthesized speech, and the excitation source provides the fine structure of the spectrum harmonics. The all-pole filter is derived from a mathematical approach that regards the speech signal as an autoregressive process, that is, the current sample is a linearly weighted sum of previous samples. This approach yields an accurate and efficient representation of the short-time spectrum of speech signals. Since the human ear is mostly sensitive to the magnitude spectrum of an acoustic signal, the ability of preserving the spectral envelope by the LP analysis is the main reason for its success.

The representation for the spectral envelope based on LP analysis also has many implications to other types of vocoders. For instance, the ceptrum, which is obtained from a homomorphic system (Figure 1–5), is an alternative form that manifests the short-time speech spectrum (Oppenheim, 1969). The impulse response, h_n, computed from the ceptrum, can be considered as the coefficient sequence of an FIR filter exhibiting a similar spectral envelope of the all-pole filter:

$$H(z) = \sum_{n=0}^{\infty} h_n z^{-n} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (1\text{–}4)$$

where a_k is the LP coefficient, and G is the gain. The filtering operation applied to the source is carried out by a convolution between the sequence, h_n, and the excitation, e_n.

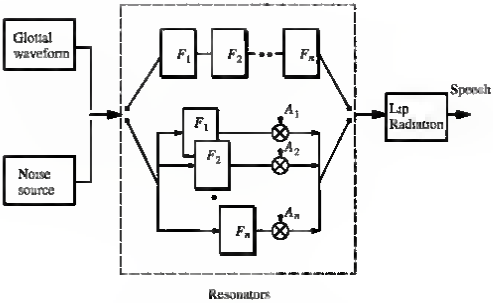


Figure 1–3 Block diagram of formant synthesizer.

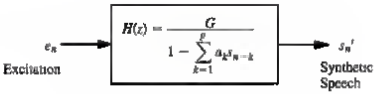


Figure 1–4. Block diagram of LP synthesizer

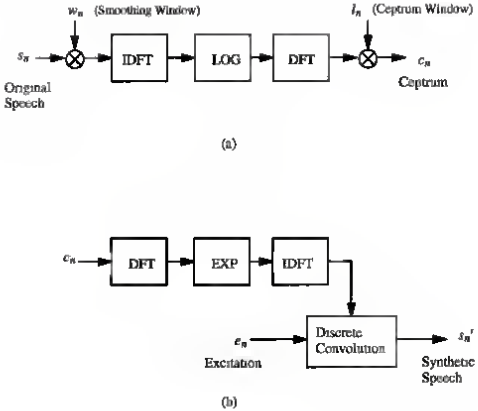


Figure 1–5 Block diagram of homomorphic system: (a) analysis process (b) synthesis process

1.3.2.4 Comments on the three types of synthesizers

The advantages and disadvantages of the three types of synthesizers are given in Table 1–1. For some LP synthesizers, the deficiency of the all-pole filter is ameliorated by employing a pole-zero model (Atal and Schroeder, 1978, Childers et al., 1981). Also the excitation function may be replaced by sophisticated pulses or innovative ensembles that simulate the residue signal. (This is discussed in detail in Section 3.1.) Moreover, an independent control of the spectral characteristics is achieved by factoring the filter into resonators and anti-resonators (Kuwabara, 1984, Childers et al., 1989b). For some formant synthesizers, the dynamics of spectral characteristics are enhanced simply by inserting features into the glottal source (Fujisaki and Ljungqvist, 1986). Likewise, the effect of source-tract interaction may be simulated by either modifying the glottal waveshape or the formant bandwidths or by incorporating a control circuit (Gufrin et al., 1976, Yea et al., 1983, Fujisaki and Ljungqvist, 1986, Wong, 1991).

Since each of the above-mentioned scheme increases the computational burden of the processing task, the complexity is no longer a major drawback only for the articulatory synthesizer. Besides, in many articulatory synthesizers the movements of the articulators are determined by comparing the formants of the synthetic speech with that of the original speech (Parthasarathy and Coker, 1992, Prado et al., 1992). Consequently, one type of source-filter synthesizer is not particularly different from the others.

1.4 Research Issues and Objectives

Ideally, a speech synthesizer should have the ability to produce any desired voice quality. From this standpoint, the attributes of voice quality are directly related to the control parameters of a synthesizer. In other words, these control parameters may verify the quality attributes that the human ear uses to discriminate voice types. This use of a speech production model for speech research is called "analysis-by-synthesis."

Table 1–1. Comments for articulatory, formant, and LP synthesizers.

	LP synthesizer	Formant synthesizer	Articulatory synthesizer
Advantage	1. few parameters are required. 2. fast algorithms are available. 3. synthetic speech is intelligible at a rate as low as 2 Kb/s.	1. glottal waveform and formant properties can be controlled independently 2. control parameters correlate with the acoustic aspects of the speech sounds. 3. source-tract interaction can be simulated by modifying the glottal source.	1. control parameters are directly related to the articulatory mechanisms. 2. source-tract interaction can be modeled. 3. articulatory parameters can be interpolated even for rapid transitions in the speech signal.
Dis-advantage	1. nasals, fricatives and stop consonants cannot be properly produced. 2. low pitch voices often sound buzzy 3. control parameters show little relation to the anatomy and physiology of speech production. 4. source-tract interaction cannot be produced in a direct manner.	1. formant/anti-formant extraction is difficult. 2. synthetic speech may sound too smooth.	1. acquisitions regarding the movements of articulators and vibratory patterns of vocal folds are difficult. 2. considerable computation is required.

Our ultimate objective in this dissertation was to develop a high-quality speech synthesizer. The quality of speech, in general, is referred to as the total auditory impression the listener experiences upon hearing the speech of a speaker. It consists of two factors, namely, naturalness and intelligibility. Through the progressive understanding of speech production researchers should be able to validate the hypothesis that the intelligibility of speech signals depends largely on the vocal tract, while the source characteristics determine the naturalness of the voice. The intelligibility is not our concern here since most present-day synthesizers are capable of conveying the intended speech content correctly. Instead, we are more interested in the vocal source because of its contributions to the naturalness of speech. For this reason, the words "quality" and "naturalness" will be considered equivalent in this dissertation.

To accomplish our objective, we decided to divide the research issues into two separate but related phases. In the first phase we discussed how to obtain acoustic measures by LP techniques. Three types of voiced speech (modal, vocal fry, breathy) were used as representative examples to illustrate the source properties.

We have selected the LP technique to accomplish this study despite some arguments against the use of such a technique. The LP analysis, in our opinion, is more than adequate because source properties are all extractable from the residue signal obtained by inverse filtering of the speech signal. This argument becomes clear in Chapter 2 when we discuss the relationship between the residue and the volume-velocity flow. We identify the significance of acoustic measures extracted from both the residue and speech signals and subsequently correlate these measures to the control parameters of a speech production model.

The knowledge gained in the first phase of the research is useful for the design of a source model for an LP synthesizer. It has long been known that the lack of glottal characteristics is the primary reason leading to the poor quality of LP synthesizers. In the second phase of the research, we first try to develop a source model to simulate the residue

signal. Such a source model will be presented in the form of a codebook that will be incorporated into a newly designed speech production model. In addition to source modeling, other factors, such as the interpolation of LP coefficients, turbulent noise, source-tract interaction, etc., have to be taken into consideration. Thus, we will present our methods and strategies to deal with these factors.

The efficacy of the source and speech production models is determined by evaluating the quality of synthetic speech. We have taken the analysis-by-synthesis approach in studying speech quality. This approach provides information about whether the important acoustic features are successfully maintained during the modeling process. While no reliable quantitative measure is available for performing speech evaluation, informal subjective listening tests were conducted to assess the quality of the synthetic speech samples. The overall research plan is presented as a schematic diagram in Figure 1–6.

1.5 Description of Chapters

Chapter 2 describes the procedures for measuring vocal source properties by linear predictive analysis. Following a retrospect of some existing acoustic measures, our first focus is on sorting the relationships between these measures and the control parameters of a comprehensive speech production model. In particular, under the guidance of this model, we propose methods for identifying and isolating the acoustic characteristics of vocal quality. Three voice types are provided as representative examples to illustrate the proposed model and analysis techniques. Knowledge gained in this chapter contributes to the understanding of general problems of source modeling and speech processing, which we present in Chapters 3 and 4.

Chapter 3 deals with the modeling of the excitation source. Depending on the voicing condition, we divide the excitation into two categories, i.e., voiced and unvoiced. A novel glottal source model is proposed to describe the voiced residue in terms of the glottal phase characteristics, while the innovation sequences are used to simulate the unvoiced residue.

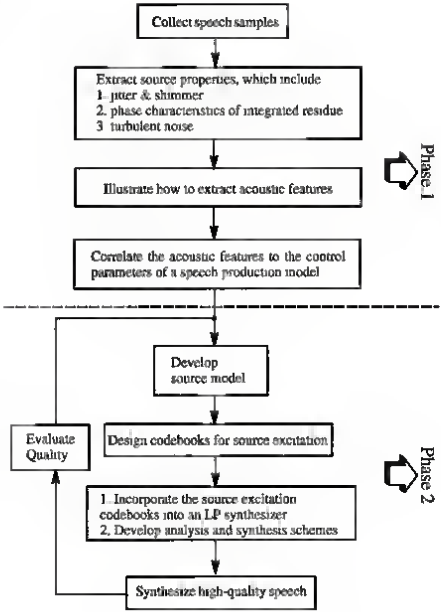


Figure 1–6. Schematic diagram of research plan.

Both types of excitations are formulated into codebooks. Our methods of generating the codebooks are associated with these two types of excitations individually

The linear predictive analysis and synthesis schemes used in this study constituted the first two parts of Chapter 4. Issues such as the voicing decision, Glottal Closure Instant (GCI) identification, codeword searching, vocal noise, source-tract interaction and gain determination are addressed. The overall performance of these schemes is dependent on how closely the reproduced speech resembles the original. While no reliable objective quality measure is currently available, we evaluate the synthetic speech by informal listening tests.

Chapter 5, the last chapter, summarizes the results of this study, discusses possible improvements to the proposed model and finally recommends some potential applications.

CHAPTER 2
SOURCE PROPERTIES

A better understanding of speech production is important for the assessment of speech quality as well as for the development of a natural-sounding speech synthesis model. In this chapter we are particularly interested in the glottal source properties that affect the perceptual quality of the voice. The elucidation of the relationship between the excitation source and the resultant speech quality requires source-related parameters to describe acoustic and perceptual features, as well as methods to extract the parameters. The analysis-by-synthesis technique is a general approach to speech analysis (Rabiner and Schafer, 1978; Furui, 1985). In principle, we establish the speech production model and then derive the model parameters used to reproduce speech signals. Speech synthesis, in conjunction with perceptual evaluation, plays a role in validating the significance of the acoustic features in terms of the model parameters. As the speech production models become more and more sophisticated, many detailed acoustic-perceptual correlations will be easily verified by the analysis-by-synthesis approach.

Our major concern is focused on efforts that will establish a relationship between model parameters and acoustic features measured from the speech signal. Following a brief review of existing acoustic measures and a background description of inverse filtering techniques, we discuss the relationship between two commonly encountered source excitation signals, namely, the residue signal and the differentiated glottal flow waveform. In order to facilitate the acquisition of the source excitation, we have used the LP technique as a vehicle to complete this research. A new LP synthesis model with appropriate source features is then proposed. Nine utterances of three types of phonations, i.e., modal, vocal

fry and breathy, were used as representative examples to validate the competence of this proposed model.

2.1 Review of Existing Acoustic Measures

Basically, researchers have used five types of acoustic measures to study vocal quality:

(1) Perturbation measures,

(2) Characteristics of the glottal flow waveform,

(3) Vocal noise,

(4) Roots of the inverse vocal tract filter,

(5) Vocal intensity

2.1.1 Perturbation Measures

Voiced speech is generated by the vibration of vocal folds. Aberrant vibratory patterns of vocal folds has long been known to result in abnormal or deviant voices (Moore, 1976). Statistical properties of the cycle-to-cycle variations in voiced speech have proven useful to characterize vocal quality (Askenfelt and Hammarberg 1986, Schoentgen 1989, Pinto and Titze 1990, Eskenazi et al. 1990). The perturbations in the fundamental frequency and amplitude of sustained utterances, termed jitter (Lieberman, 1961) and shimmer (Koike, 1969), respectively, were two of the first acoustic measures reported to be correlated with vocal pathology. Since then, other perturbation measures have also been shown to be capable of distinguishing pathological from normal voices.

2.1.2 Characteristics of the Glottal Flow Waveform

The characteristics of the glottal flow considered for the assessment of speech quality can be further classified into two categories. (1) qualitative analysis based on parameters of source models, and (2) spectral tilt.

2.1.2.1 Quantitative analysis based on parameters of source models

Monitoring the glottal flow waveform is a direct means of studying the variations of the glottal source (Hillman and Weinberg, 1981, Javkin et al., 1987; Price, 1989). In order to assess such variations on a quantitative basis, a parametric model was often introduced. One such model that has been widely adopted for quality assessment in recent years is the LF model (Fant et al., 1985, Fujisaki and Ljungqvist, 1986; Fant and Lin, 1988; Gobl, 1988 & 1989, Karlsson, 1988, Ahn, 1991, Tenpaku and Hirahara, 1990, Childers and Lee, 1991). This model is useful because it ensures an overall fit to commonly encountered differential glottal pulses with a minimum number of parameters, and it is flexible in the extent to which it can match various phonations.

2.1.2.2 Spectral tilt

In addition to the parametric variations of the source models, the spectral tilt of the glottal flow appears to be characteristic of different voice types (Holmes, 1974; Hiki et al., 1976, Monsen and Engebretson, 1977). In fact, the steepness of the spectral tilt is caused by the rapidity of the closing phase and by the abruptness of the glottal closure. The perceived quality of speech is related to the spectral tilt (Childers and Lee, 1991). A steeply declining spectral tilt results in a lax quality, whereas a gradually declining tilt produces a tense quality. To achieve a quantitative measure of this aspect of vocal quality, the spectrum of the glottal flow is usually approximated by a three-pole model, or equivalently a two-pole model for the differentiated glottal flow. The coefficients of the three- or two-pole models are then used to indicate the spectral tilt.

2.1.3 Vocal Noise

Turbulence at the level of the glottis also contributes vocal quality such as hoarseness and breathiness, which is a prominent symptom of laryngeal pathologies (Klatt, 1987, Klatt and Klatt, 1990; Childers and Lee, 1991). Methods for measuring the turbulent noise consist

of the relative intensity (Hiraoka et al., 1984, Fukazawa et al., 1988), the spectral noise level and the harmonic-to-noise ratio (Katajima, 1981; Yumoto et al., 1982, Yumoto et al , 1984, Kasuya et al., 1986 a&b, Muta et al., 1987, Childers and Lee, 1991) In most cases, these noise measures were influenced by the spectral content of the analyzed speech Consequently, better methods are needed so that the glottal flow waveform can be analyzed more precisely.

2.1.4 Roots of the Inverse Vocal Tract Filter

Another aspect of speech spectra that affects the detection of laryngeal dysfunction has been demonstrated by Deller and Anderson (1980), who represented the speech signal by the roots of the inverse filter and then applied pattern recognition techniques to dichotomize the subjects as either normal or pathological. It was found that the discrimination function employed in detecting laryngeal behavior was more sensitive to the poles attributable to the glottal source than to the formant structure (Deller, 1982) This technique was later applied to the EGG signal by Smith and Childers (1983) as a method for detecting laryngeal pathology They concluded that the LP features of EGG signals were more sensitive to pathology detection than similar parameters measured from speech signals. Recently, the same task was recast on the pattern analysis of LP coefficients by vector quantization (Childers and Bae, 1992). Inferences based upon their results were consistent with previous research.

2.1.5 Vocal Intensity

Vocal intensity is less specific in quality assessment (Colton 1973 Hollien 1974) It is largely irrelevant to the perceived quality except for loudness. Since the selected speech samples we used were approximately at the same power level after digitization, vocal intensity was not considered an important factor in our research

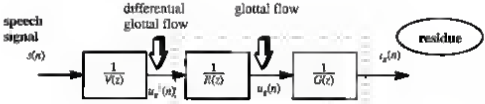
2.1.6 Remark

A multivariate statistical analysis of acoustic parameters may result in a quality predictor that matches well with the objective evaluation (Hiki et al , 1976, Wolfe and Steinfatt, 1987 Eskenazi et al. 1990, Pinto and Titze, 1990) In addition, measures of higher orders may also provide extra degrees of freedom in statistical analysis. Using these measures in quality assessment causes difficulties in justifying the significance of each individual measure and their correlations. Pinto and Titze (1990) made an attempt to unify existing jitter, shimmer and noise measures. however, no effort was made to sort out the relation between the acoustic measures and the control parameters of a specific speech production model This motivated us to explore those relations.

2.2 Glottal Inverse Filtering

Glottal inverse filtering is a popular and efficient means for investigating the acovities of the glottal source. It is based on the assumptions that the source excitation and the supraglottal loading are separable and that the source properties of the speech production model can be uniquely determined. The principle of inverse filtering is to obtain the glottal flow by eliminating the effects of vocal tract transfer function and lip radiation from the speech signal Figure 2--1 presents the conceptual inverse filtering model. Noooe that in this representation the sequence of the vocal tract transfer function and lip radiation are reversed because the speech production is assumed to be a linear model

Current methods for glottal inverse filtering center on LP analysis (Berouti, 1976, Wong et al., 1979, Matausek and Batalov, 1980. Childers and Lacar, 1984; Krishnamurthy and Childers, 1986, Milenkovic, 1986, Childers and Lee, 1991) Among the various methods the closed phase covariance analysis is considered the most reliable because no source-tract interaction is involved. However, the disadvantages of this method are (1) it needs to locate the closed phase very accurately, and (2) it is only feasible when the close phase is long enough to accommodate the analysis window



G(z) : glottal shaping filter
V(z) : vocal tract transfer function
R(z) : lip radiation

Figure 2–1 Block diagram of glottal inverse filtering.

Recently, in order to alleviate these disadvantages, adaptive approaches have been used to track the rapid change of the parameters of the vocal tract during the glottal closed phase (Ting and Childers, 1990). In fact, it is more convenient to estimate the composite effect of the glottal pulse, lip radiation and vocal tract together. The vocal tract transfer function could be obtained by removing the source-related roots from the LP polynomial (Childers and Lee, 1991). However, this approach may introduce errors due to the incorrect elimination or merging of such roots. Furthermore, since the estimate of the vocal tract parameters is based on an entire pitch period, the effects of different damping factors caused by the open and closed glottal intervals during a pitch period affect the estimate. Consequently, the estimated glottal flow waveform becomes an "average" waveform for the entire pitch period. This average waveform may not be truly representative of the actual waveform.

From the discussion above, we know that the glottal flow waveform is not always obtainable using the glottal inverse filtering techniques. However, the estimation of residue signal is seldom affected by the preceding factors. Moreover, as will be seen in the next section, the retrieval of the glottal phase characteristics can be resolved from the residue signal. For these two reasons, we concentrate our study on the residue signal. The potential of the residue can be seen from its appearance. It has been observed that the residue extracted from normal voices consists of periodic sharp spikes and low-level noise components, whereas the residue extracted from deviant voices exhibits a less distinctive pattern of periodic spikes (Figure 2–2). Because such an observation is not as noticeable as in the speech signal, many researchers advocated the use of the residue signal over the speech signal for the analysis of abnormal voices (Koike and Markel, 1975, Sorensen and Horn, 1984, Prosek et al., 1987). Ironically, the quantitative measures deduced from the residue signal failed to support their claims (Schoentgen, 1982). We believe this contradiction is due to the inadequacy of the acoustic measures and the analysis methods. It was noted that the LP coefficients calculated by a fixed frame autocorrelation method, which was used by

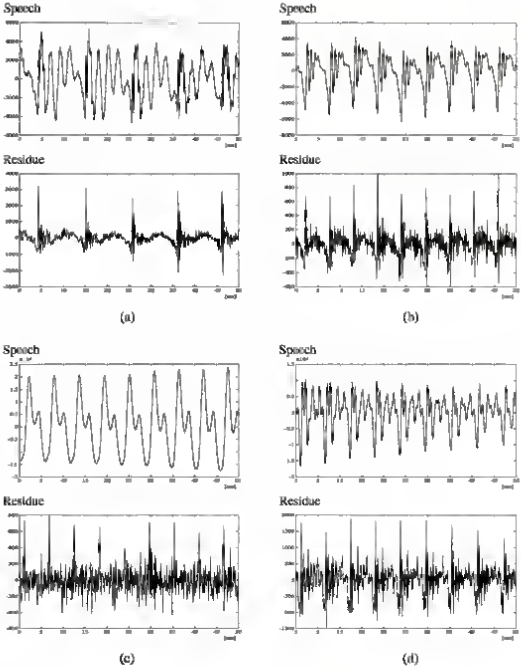


Figure 2–2 Speech and residue waveforms for two normal subjects (a) and (b), and for two pathological subjects (c) and (d). The pathological symptom is hoarseness for subject (c) and is bilateral paralysis of TVC for subject (d)

Schoentgen (1982), were affected by the size and position of the analyzed frame. Any small deviation of the estimated coefficients could result in a great change of the residue signal (Ananthapadmanabha and Yegnanarayana, 1979). Consequently, the acoustic measures derived from the fixed-frame autocorrelation method are prone to error. To avoid this problem, a pitch-synchronous covariance analysis method has been used (Chandra and Lin, 1974).

2.3 Correlation between Residue and Differentiated Glottal Flow

It is constructive for us to clarify the relation between the residue and the glottal flow before we explore the characteristics of the glottal source. As shown in Figure 2–1, the inverse filtering can be imaged as a process of unscrambling the speech signal so as to obtain the excitation waveform. One of the intermediate products is the differentiated glottal flow, while for our purpose the residue signal is the ultimate result. Thus, the correspondence between the residue and glottal flow can easily be illustrated as a filtering process. Here we adopt a two-pole filter to model the spectrum of the differentiated glottal flow. The filter coefficients are obtained by an LP analysis of the modeled LF waveform.

Since the LF model has been successfully used to describe the characteristics of a differentiated glottal flow, we adopt it as an explanatory media for the subsequent discussion. The equations of the LP-model are given as

$$E(t) = \begin{cases} E_0 e^{\alpha t} \sin \omega_g t & 0 \le t < t_e \qquad \text{(2–1a)} \\ -\dfrac{E_e}{\xi t_a}\left[e^{-\xi(t-t_e)} - e^{-\xi(t_c-t_e)}\right] & t_e \le t \le t_c \qquad \text{(2–1b)} \end{cases}$$

where t_p, t_e, t_c are parameters related to the glottal flow peak, maximum closing rate and glottal closure, respectively. The parameter t_a is used to control the abruptness of return phase, and the parameter ω_g, defined as π/t_p, determines the frequency of sinusoid. Parameters E_0, α and ε are for computational use only. A typical LF-model waveform is shown in Figure 2–3.

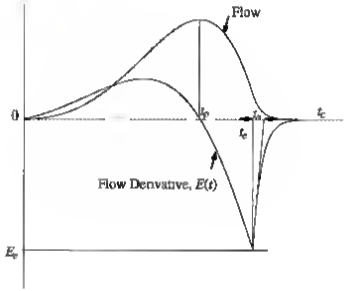


Figure 2–3. LF-model waveform, $E(t)$ for the differentiated glottal flow

The first segment of the LF model characterizes the differentiated glottal flow over the interval from the glottal opening to the maximum negative excursion of the waveform. The second segment represents a residual glottal flow that comes after the maximum negative excursion. It can be shown from Eq (2–1) that the spectrum of the first segment is dominated by the exponential component, $e^{\alpha t}$, of which the "negative bandwidth" equals α/π. Likewise, the frequency response of the second segment can be approximated by a first order lowpass filter with a cutoff frequency $F_c = 1/(2\pi t_a)$ (Fant and Lin 1988). As a result, the bandwidths of the first and second segments, B_1 and B_2, are

$$B_1 \approx \frac{\alpha}{\pi} \tag{2–2}$$

$$B_2 \approx \frac{1}{2\pi t_a} \tag{2–3}$$

It can be shown that the poles of the filter are either both real or a complex conjugate pair. The center frequency ω and bandwidth B can be calculated from the zeros, z_i's, by

$$\omega = \tan^{-1}\left[\frac{im(z_i)}{re(z_i)}\right], \tag{2–4}$$

$$B = -\ln|z_i| \tag{2–5}$$

We have found that the center frequency ω of the poles and ω_g of the LF model are nearly the same. Thus we are only concerned with the change in bandwidths of the poles of the inverse filter. The bandwidth of source spectrum B is, in general, very close to B_1, causing the waveshape of the first segment to be obliterated after inverse filtering. However, B_2 is much higher than B. The second segment thereby retains its waveshape after inverse filtering although the resultant phase may be different from the original. A typical example is given in Figure 2–4, which displays the spectra of the first and second segments of LF-model as well as the corresponding spectrum of the two-pole model. As a result, the residue derived from the LF model waveform has a flat spectrum envelope and exhibits a sharp pulse at the conjunction between two segments, where the glottal closure occurs in the LF model.

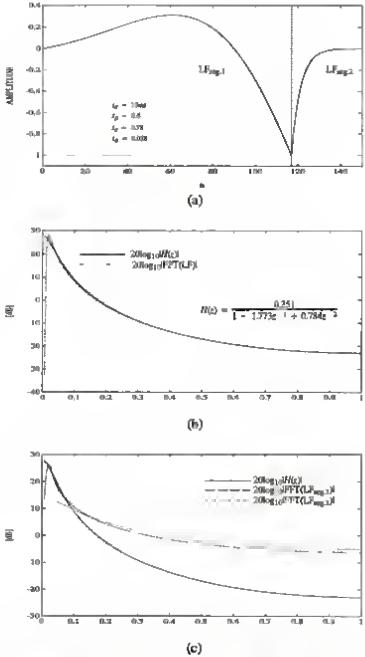


Figure 2–4 Effects of the inverse filter imposed on the differentiated glottal waveform: (a) LF-model waveform, (b) FFT spectra of LF model waveform and two-pole model $H(z)$, (c) FFT spectra of individual segments of the LF model.

Knowing the relationship and transformation between the differentiated glottal and residue signals should enable us to retrieve one signal from the other. Although the residue signal does not appear to be highly informative, its integral, in contrast, tends to partially re-exhibit the shape of the first segment of the differentiated glottal flow. Thus, the analysis strategies based on the differentiated glottal flow can be transplanted to the integrated residue with little modification. To support our claim, we perform the inverse filtering based on a synthetic vowel, as shown in Figure 2–5, so that the similarities and contrasts between the LF-model waveform and the integrated residue can be noticed readily.

2.4 Choice of Model Type

In essence, building a speech model is equivalent to systematically coordinating the acoustic and perceptual attributes into a joint construct. For speech synthesis, modeling is usually aimed at the parameterization of the voice source and the vocal tract.

In Chapter 1, we have shown that the speech production and propagation mechanisms can be described by a source-filter model (Fant, 1960), consisting of the glottal source, vocal tract transfer function, and lip radiation. This model is not only simple but effective in characterizing speech signals. The formant and LP synthesizers belong to this group, and both synthesizers were used to study vocal variability (Rosenberg, 1971; Holmes, 1973; Sambur, et al., 1978; Atal and David, 1979; Kuwabara, 1984; Hermansky et al., 1985; Hedelin, 1986; Klatt, 1987; Muta et al., 1987; Childers et al., 1989b; Childers and Wu, 1990; Klatt and Klatt, 1990; Childers and Lee, 1991; Lalwani and Childers, 1991).

For the formant synthesizers, the properties of each component of the source-filter model are elaborated individually. Usually, the lip radiation is approximated by a differentiator. The vocal tract transfer function is characterized by the formants/anti-formants, which are implemented by resonator/anti-resonator filters. The source model is designed to imitate the glottal volume velocity waveform. Speech quality generated from

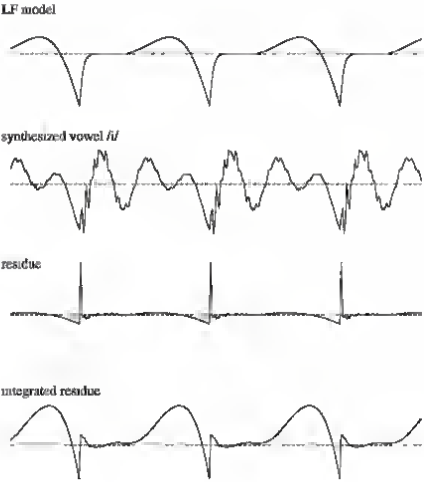


Figure 2–5 Illustration of the similarity between the differential glottal flow and the integrated residue signal. Waveforms from top to bottom are (1) LF model, (2) synthesized vowel */i/* produced by the LF model, (3) residue signal, and (4) integral of the residue signal.

such synthesizers was judged to be satisfactorily high, provided that the glottal flow is appropriately modeled (Klatt, 1980, Holmes, 1983, Pinto et al., 1989)

For the LP synthesizers, the composite frequency response of the glottal flow, vocal tract and lip radiation is modeled by a slowly time-varying filter (Atal and Hanauer, 1971) The associated source excitation is primarily used to account for the periodicity of the glottal pulse, in other words the pitch The synthetic speech quality for many previous models is considered unnatural due to an oversimplified excitation, failures to properly identify voicing, and poor spectral resolution (Woog, 1980, Kahn and Garst, 1983) However, with the use of sophisticated source excitations, the LP synthesizer can still achieve a very high quality

On the whole, it appears that the perceptual quality of synthetic speech is improved by improving source excitation models for both LP and formant synthesizers. In the past the lack of a physiological interpretation for the residue was considered the primary obstacle against the use of LP techniques in examining a specific vocal quality, making the formant synthesizer a more popular tool Nonetheless, this argument may no longer be valid once we are able to verify the relationship between the residue and the glottal flow Since the LP synthesizer has the advantages of: (1) computational efficiency, and (2) ease of obtaining the residue from speech, we decided to use the LP synthesis model as the means to accomplish this research

We start by integrating the acoustic attributes into a comprehensive model Our strategies in constructing a high-quality LP speech production model follow the analysis-by-synthesis rules. This speech production model is depicted in Figure 2–6, while Figure 2–7 presents the correlations we are going to examine between the acoustic parameters and model parameters. Before we work out the details, there is much groundwork to establish

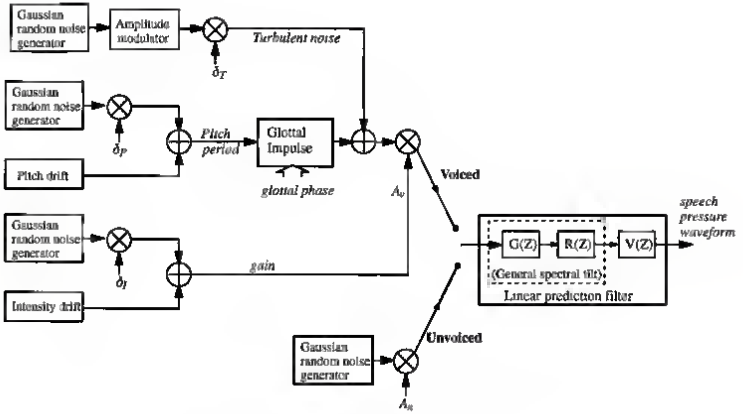


Figure 2–6. A comprehensive linear prediction speech production model.

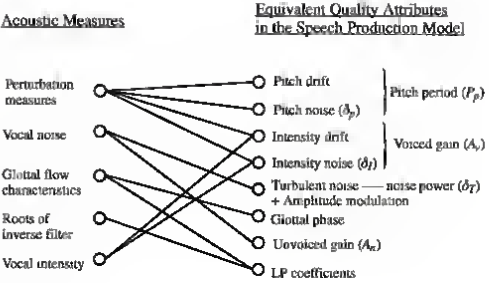


Figure 2–7 Correlations between acoustic attributes and model parameters: the notation "A — B" is read as "A is related to B."

2.5 Data Collection and Methodological Consideration

In addition to the description of the experimental data base, this section provides background information about vocal quality including modal vocal fry, and breathy voices. The source properties quoted from previous research are listed for the purpose of comparing our analysis results. After explaining the measuring methods, we discuss the pre-processing schemes required to extract the acoustic features. These preparations constitute the foundation for the source extraction.

2.5.1 Experimental Data Base

The vowel */a/* was chosen to this experiment because it has been found useful for ultra high speed laryngeal photography. Nine utterances for three different voice types, i.e., modal, vocal fry, and breathy, served as our data base. All these utterances were categorized by professional speech scientists. A description of the data base is shown in Table 2–1.

Table 2–1. Data base for speech analysis.

Subject	Sex	Voice type	# of pitch periods
M1	M	model	382
M2	M	model	331
M3	M	model	244
V1	M	vocal fry	176
V2	M	vocal fry	239
V3	M	vocal fry	109
B1	M	breathy	201
B2	M	breathy	273
B3	M	breathy	454

During speech processing, the measures were performed over a steady state interval. As will be discussed later, the acoustic measures required a precise identification of the pitch period. An additional signal, the electroglottograph (EGG), was employed in this study to aid the speech processing. We sampled the speech and EGG signals at 10 KHz with 16-bits

precision Both signals were digitized simultaneously using Digital Sound Corp DSC–240 preamphfier and a DSC–200 digitizers. The microphone was an Electro–Voice RE–10 held six inches from the lips. Before digitization the signals were bandlimited to 5 kHz by anti-aliasing, passive, elliptic filters with a minimum stopband attenuation of –55dB and a passband ripple of ±0·2 dB All data recordings were collected in an Industrial Acoustics Company (IAC) single wall sound booth To compensate for the microphone characteristics at low frequencies, the frequency response of the speech recordings was further corrected using a linear phase FIR filter

2.5.2 Vocal Quality

The adequacy of an acoustic measure can be illustrated by its capability for characterizing vocal quality. When assessing the acoustic measures, we certainly need to have a general concept of the vocal quality. As mentioned in the previous chapter, the vocal quality is referred to as the auditory impression the listener experiences upon hearing the speech of another talker. Major types of vocal quality, according to Laver and Hanson (1981), are modal, breathy, vocal fry, falsetto, harshness, and whisper. We excluded falsetto, harshness and whisper from this study because the other three voice types were considered sufficiently representative of three modes of vocal fold vibratory patterns. The qualitative definitions (Lieberman and Blumstein, 1988, Eskenazi et al , 1990) of the three voice types are

Modal : Defined as a normal phonation. A modal phonation is characterized by a moderate frequency, wide lateral excursions, and complete closure of the glottis during about one third of the entire pitch period.

Breathy : Defined as audible escapage of air through the glottis due to insufficient glottal closure. The degree of breathiness severity is inversely proportional to the length of the closed glottal phase.

Vocal fry : Defined as a low-pitched, creaky kind of phonation. It also shows a great deal of irregularity from one pitch period to the next.

In this study, we are interested more in the source characteristics than in the vibratory frequency of the laryngeal vibration. This is because the effect of the glottal vibration in terms of vocal registers is already reflected in the categorization of various voice types. Some acoustic features of glottal factors of various voice types are summarized in Table 2–2. These features will serve as references when we examine speech features using proposed acoustic measures.

Table 2–2. Summary of acoustic characteristics of glottal sources for three voice types.

		Modal	Vocal fry	Breathy
Fundamental frequency		medium	low	medium
Perturbation measure	Jitter	low	high	high
	Shimmer	low	low	high
Properties of glottal flow	Turbulent noise	medium	low	high
	Pulse width	medium	short	long
	Pulse skewness	medium	high	low
	Abruptness of closure	medium	fast	slow
Spectral tilt		medium	flat	steep
Vocal intensity		wide range	low	low

Source. Ahn, 1991, Childers and Lee, 1991

2.5.3 Analytical Logics

In most research the characteristics of glottal flow for one pitch period are delineated using variables consisting of either the relative timing or the durations of special events such as the glottal opening and closure. Because the pitch period is usually a known value, it is

the waveshape, rather than the absolute timing, that has attracted the researchers' attention. This suggests a standardization procedure for those variables based on the underlying pitch period. Properties of the standardized variables drawn from a large population are assumed to represent general characteristics of the glottal source. Many postulates and conclusions pertaining to vocal quality are thereby deduced based on the statistical results. Alternatively, such a statistical analysis can be performed by evaluating the averaged glottal pulse over a large number of sample periods. Such a logical variant will facilitate the inquiry of some timing events in the glottal flow, the differentiated glottal flow and the integrated residue.

2.5.4 Standardization of Pitch Period

To perform the alternative statistical analysis suggested above, we resample every pitch period at a variable rate so that every digitized waveform has the same length. In other words, the sampling rate for each individual pitch period should be different in order to make the digitized waveforms summable. Difficulties associated with this procedure lie in the identification and standardization of each pitch period. A direct and exact solution, from a mathematical point of view, is the Sinc-interpolated sampling rate conversion (Schafer and Rabiner, 1973, Kroon and Atal, 1990; Schumacher and Chafe, 1990). The Sinc interpolation is given by

$$\hat{x}(nT') = \sum_{i=-\infty}^{\infty} x(i) \frac{\sin\left\{\pi f_s[nT' - \frac{i}{f_s} + \theta]\right\}}{\pi f_s[nT' - \frac{i}{f_s} + \theta]} \qquad (2\text{–}6)$$

where f_s is the sampling frequency of the original sequence $x(i)$, T' is a new sampling interval of $\hat{x}(n)$, and θ is a phase offset. The T' and θ for each individual period are determined so as to yield a maximum similarity across the resampled periods.

The processing task required by Sinc interpolation is computationally expensive. A simpler approach is presented below to facilitate the computation. First of all, we interpolate

the analyzed signal $s(n)$ by a factor of five times by using a lowpass filter

$$s(n) = s_i(n) * \hat{h}(n) \tag{2–7}$$

where [*] denote the convolution, $s_i(n)$ is the linearly interpolated data sequence, and $\hat{h}(n)$ is the impulse response of a lowpass filter with the cut off frequency at $\pi/5$. In our case, a 511-order FIR filter designed by using the window method is employed to avoid phase distortion. The impulse response of this FIR filter is

$$\hat{h}(n) = \frac{1}{K} h(n) w(n) \tag{2–8}$$

where
$$h(n) = \frac{\sin(\pi n \frac{1}{5})}{\pi n \frac{1}{5}} \qquad \text{for } n = 0, \pm 1, \pm 2, \ldots; \tag{2–9}$$

$$w(n) = \begin{cases} 54 - .46\cos(\frac{2\pi n}{255}), & |n| \le 255 \\ 0, & \text{otherwise,} \end{cases} \tag{2–10}$$

$$K = \sum_{n=-255}^{255} h(n) w(n) \tag{2–11}$$

The next step is to separate each individual pitch period along the signal. We use the two-channel approach (Krishnamurthy and Childers, 1986) to assist this processing automatically. The glottal closure instant, which is signaled by a rapid decrease in the EGG, has been found to coincide with the minimum in the differentiated EGG (DEGG) for that period. Thus, we can locate the instant of glottal closure by picking the negative peaks of the DEGG signal, as illustrated in Figure 2–8. A pitch period is then defined as the interval between two consecutive glottal closure instants.

Due to the propagation delay of the sound wave from the glottis to the microphone, we apply a time lag of 0.9 msec to the EGG signal to achieve synchronization with the speech signal. Also, in order to improve the accuracy of the locations of the peaks, we employ a quadratic interpolation method (Markel and Gray, 1976, Titze et al., 1987). Let $f(-1)$, $f(0)$,

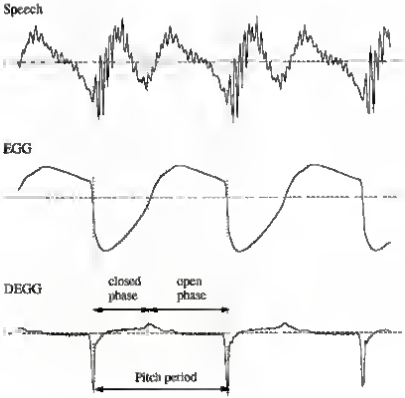


Figure 2–8: Synchronized speech, EGG, DEGG signals.

and $f(1)$ define three points centered at a peak, where $f(0)$ corresponds to a discrete minimum value, and $f(-1)$ and $f(1)$ are points to the left and right of $f(0)$. The position of interpolated minimum, λ, is then resolved from a second-order approximation among three points by

$$\lambda = -\frac{f'(0)}{f''(0)} = -\frac{f(1) - f(-1)}{2f(1) - 2f(0) + f(-1))} \qquad (2\text{–}12)$$

Values obtained through this process are rounded to the nearest sample of the re-sampled signal. Thus, the resulting resolution of each pitch period, estimated from the re-sampled sequence, increases by approximately five times.

Finally, the length of the pitch period is adjusted to 512 samples by using the FFT method. That is, depending on the number of the samples in the interpolated period, we append zeros or remove the high frequency range of the FFT sequence to achieve the intended length (512 samples in this case). The fixed-length signal is then obtained by taking the IFFT of the resultant FFT sequence. Notice that the discontinuity (linear trend) between two boundaries of the underlying signal must be removed before applying the FFT method since this signal has to be circularly periodical.

2.6 Feature Extraction

Under the guidance of the proposed speech production model (Figure 2–6), we explore the relation between the modal parameters and some existing acoustic measures. A pitch-synchronous covariance LP analysis is adopted to estimate the spectral properties of the speech signal. The LP order is chosen to be 14 to account for the spectral tilt of the glottal flow and the number of formants within 5 KHz bandwidth. Following the arrangement presented in the survey of acoustic measures, we illustrate how to extract model parameters that correspond to the perturbation measures, spectral tilt, phase characteristics, and vocal noise sequentially. The examination regarding the roots of the inverse filter is not considered in this study.

2.6.1 Perturbation Measure

As mentioned previously, the perturbation measures are used to characterize the vibratory patterns of the vocal folds, which include variabilities in the pitch period and waveform amplitude. The perturbation measures can be further divided into two types, namely, subharmonics and random noise as demonstrated in Figure 2–9. The subharmonics result from a repetitive vibratory pattern extending more than one pitch period, while the random noise represents the unpredictable characteristics of the vocal fold vibration. To avoid further complicating the problem, we confined our research to random noise while illustrating the proposed perturbation measures.

Because each subject involved in this experiment was instructed to utter a steady vowel with a comfortable intensity, the pitch and intensity contours of recorded speech were considered to be fairly stable. Typical pitch and intensity contours of a modal voice are shown in Figure 2–10(a) and Figure 2–11(a). As we are interested in the perturbation associated with the measured signal, a proper initial step is to obtain the corresponding deviation by removing the average value. By inspecting the spectral properties of the deviations of pitch and intensity signals (Figures 2–10(b) and 2–11(b)), we find that both spectra are relatively flat except at the region of low frequencies. This finding leads us to conjecture that the deviation signal can be modeled as a slow fluctuating component accompanied with a white noise source. The low-frequency component in the deviation signal, termed "drift" in many studies, is known as the inherent nature of human speech. Though the dynamic patterns of drift determine the tune in speech, they are unlikely to provide much information about vocal quality. Thus, we are safe in discarding this effect while studying vocal quality. In fact, as pictured from the point of view of the filtering process, most perturbation measures were introduced to eliminate the drift or to emphasize the white noise source. Examples for some perturbation measures and their mathematical relationships can be seen in Pinto and Titze (1990).

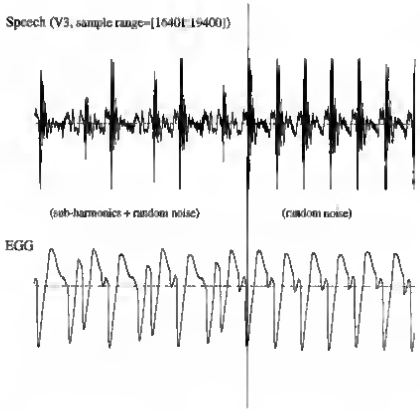


Figure 2–9. Demonstration of two types of perturbations: subharmonics and random noise.

The use of a highpass filter will not properly separate the noise source from the drift because it removes the low-frequency portion of the noise also. Thus, we performed the separation in the frequency domain using a DFT method with the following steps:

1. We remove the linear trend of the two end boundaries of the deviation signal to avoid large discontinuities occurring at boundaries due to the DFT method.
2. Before computing the DFT sequence of the resultant deviation signal, we further remove the d.c. component introduced by the first step.
3. Except the d.c. component, the magnitude of the DFT sequence below one-fifth sampling rate ($\pi/5$) is set as the average magnitude of the rest DFT sequence (see Figures 2–10(b) and 2–11(b)).
4. Given the new DFT sequence with the phase unchanged, we then take the inverse DFT of the new sequence to yield the noise signals (see Figures 2–10(c) and 2–11(c)).

Examples of the histograms of the noise signals are shown in Figure 2–12. It appears that the zero-mean Gaussian distribution provides a good fit for the underlying frequency distribution. We therefore assume that the noise component exhibits a Gaussian distribution, in which the standard deviation is sufficient to characterize the statistical property. This hypothesis can be informally validated by inspecting the cumulative probability density functions of the noise components and by comparing them with the corresponding Gaussian distribution function with the same mean and variance (Figure 2–13).

Following the terminologies defined by Pinto and Titze (1990), we use δ_P and δ_I to denote the standard deviations of the pitch and intensity noise components respectively. The following discussion illustrates how the δ_P and δ_I are related to the jitter and shimmer. We define the normalized jitter (in percent) as

$$\% \text{ jitter} = \frac{1}{n}\sum_{i=1}^{n}\frac{|P_i - P_{i-1}|}{P_n} \times 100\% \qquad (2\text{–}13)$$

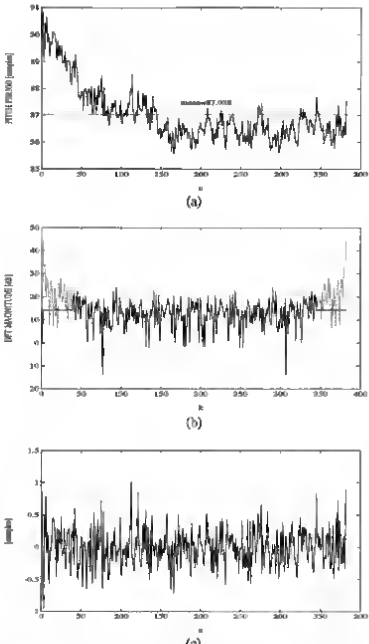


Figure 2–10. Extraction of pitch noise (a) original pitch contour (b) magnitude FFT (dotted line) of the deviation signal and the one after adjustment (solid line) (c) pitch noise after taking the inverse DFT of the adjusted DFT sequence.

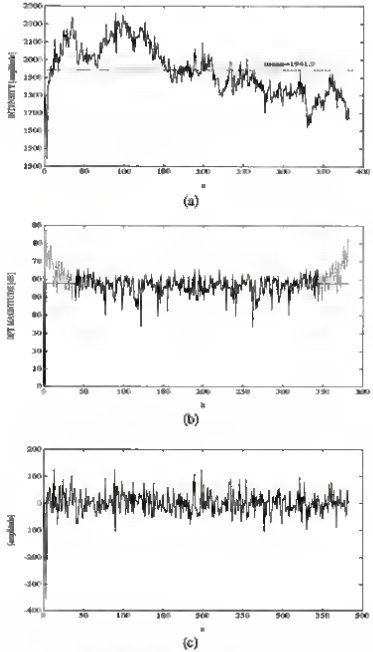
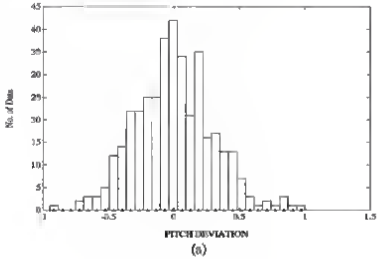


Figure 2–11 Extraction of intensity noise (a) intensity contour (b) and (c) are the same as in Figure 2–10



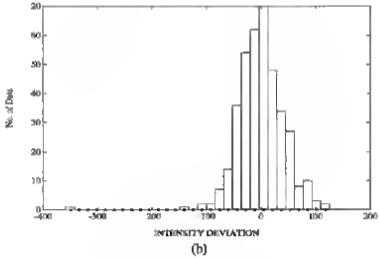
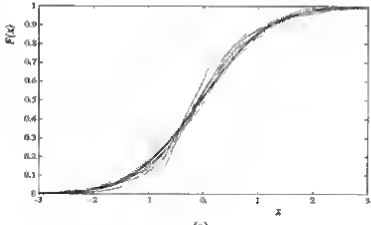
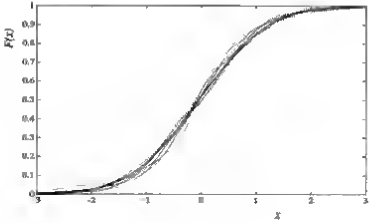


Figure 2–12. Histograms of (a) pitch noise, and (b) intensity noise.



(a)



(b)

Figure 2–13 Normalized cumulative probability distribution functions ($F(x)$'s) of perturbation noises for nine utterances: (a) pitch noise, (b) intensity noise (shown in dotted lines). The corresponding Gaussian distribution with the same variance is drawn by the solid line

where |•| denotes the absolute value, P_i is the ith pitch period in a segment of n pitch periods and P_0 is the mean value of P_i's. If we define $P_d^i = P^i - P_0$ then Eq (2–13) can be rewritten as

$$\% \text{ jitter} = \frac{1}{n}\sum_{i=1}^{n} \frac{| P_0 + P_d^i - (P_0 + P_d^{i-1}) |}{P_0} \times 100\%$$

$$= \frac{1}{P_0}\sum_{i=1}^{n} \frac{| P_d^i - P_d^{i-1} |}{n} \times 100\% \qquad (2\text{–}14)$$

By assuming P_d to be a random process with a zero-mean Gaussian distribution and $n \gg 1$, the equation can be approximated by

$$\% \text{ jitter} \approx \frac{\sqrt{2}}{P_0} \overline{| P_d^i |} \times 100\%$$

$$\approx \frac{\sqrt{2}}{P_0}\sqrt{\frac{2}{\pi}}\,\delta_P \times 100\% = \frac{2}{\sqrt{\pi}}\delta_P \frac{100\%}{P_0} \qquad (2\text{–}15)$$

where the overbar denotes the statistical mean. In a similar manner, the percent shimmer can be derived as

$$\% \text{ shimmer} = \frac{1}{n}\sum_{i=1}^{n} \frac{|A_i - A_{i-1}|}{A_0} \times 100\%$$

$$\approx \frac{2}{\sqrt{\pi}}\delta_I \frac{100\ \%}{A_0} \qquad (2\text{–}16)$$

where A_i defines the square root of the intensity (rms power) of the ith glottal period, and A_0 is the average of A_i's. Compared to the definition given by other researchers, where A_i is the peak magnitude, the adopted form is more likely to correspond with the perceptual characteristics of the human auditory apparatus that resolve short-time spectra of acoustic signals. More important, it is the power density rather than the peak magnitude used for the speech analysis and synthesis in the proposed speech production model.

From Eqs (2–15) and (2–16) we know that the %jitter and %shimmer are just other types of manifestations for the pitch and intensity noise. To verify the foregoing derivation, we list the δ_p's and δ_I's measured from the pitch and intensity noise signals as well as those derived from %jitter and %shimmer in Table 2–3. The tabulated values are very close to the values computed from two different approaches; such results thereby substantiate our assumptions with regard to the perturbation noise and its relation to other perturbation measures.

Table 2–3. Mean values and standard deviations ($\pm$ std) of the acoustic measures for each subject.

	Pitch period [us]	%jitter	δ_p (estimated)	δ_p	A_I (normalized)	%shimmer	δ_I (estimated)	δ_I
M1	8 703 $\pm$ 0 103	0 381	0 294	0 299	1.000 $\pm$ 0 072	2 352	40 475	43 822
M2	7 731 $\pm$ 0 049	0 317	0 217	0.232	1 000 $\pm$ 0 110	3 781	27.341	24 903
M3	7 709 $\pm$ 0 062	0 344	0 235	0.249	1 000 $\pm$ 0 053	2 763	80 154	72 207
V1	11 105 $\pm$ 0 099	0 566	0.557	0.515	1 000 $\pm$ 0 053	2.428	44.724	42.325
V2	7.550 $\pm$ 0 136	0 474	0.317	0 297	1 000 $\pm$ 0 233	2.751	44.389	41 408
V3	24 993 $\pm$ 4 725	10.005	22.162	32.663	1 000 $\pm$ 0 164	9 010	543.521	619.006
B1	9 254 $\pm$ 0 105	0.893	0.733	0.746	1 000 $\pm$ 0 053	2.733	220.252	213.224
B2	8 908 $\pm$ 0 222	0 803	0.634	0 789	1 000 $\pm$ 0 166	11 140	999 290	1089 19
B3	4 697 $\pm$ 0 042	0 475	0.198	0 177	1 000 $\pm$ 0 084	0.847	91 759	86 989

2.6.2 Spectral Tilt

Theoretically, the transfer function of the vocal tract is characterized by a set of formants, which are distributed along the frequency axis. The spectral tilt of speech is mostly dominated by the lip radiation and glottal shaping filter (Rabiner and Schafer, 1978). If the lip radiation is modeled as a differentiator, then the differenoated glottal flow becomes the most pertinent component to determine the spectral tilt of a speech signal. This implies that the spectral tilt of the differentiated glottal pulse can also be estimated from the speech signal.

As discussed in Section 2.3, we used a two-pole filter to approximate the spectral tilt of differentiated glottal flow. The filter coefficients are now estimated using LP analysis based on the speech signal. Table 2–5 lists the estimated LP coefficients for the three different voice types.

2.6.3 Glottal Phase Characteristics

In addition to a general comparison of the glottal phase characteristics for the nine utterances, we present a novel measure called "abruptness index" to depict the return phase of the glottal flow.

2.6.3.1 General properties

Because the magnitude spectrum of the residue signal is flat (not in a strict sense if we consider the spectral harmonics and modeling errors) due to the inverse filtering, phase characteristics are the only information left in the residue that is related to the glottal source (Wong and Markel, 1978; Hedelin, 1988). As mentioned earlier, it is the integrated residue which resembles the differentiated glottal flow, exhibiting certain physiological features of the vocal folds. Presumably, we can explore the phase properties of the glottal source by examining the integrated residue. Unfortunately, the timing of transitional glottal events are distorted by the inverse filtering and integration. Timing factors extracted from the integrated residue are not as useful as those for the glottal flow and, therefore, will be not

investigated. Instead, the comparison across the integrated residues of the nine utterances is performed using correlation coefficients. The results are given in Table 2–4. These results, however, do not adequately characterize the correlation between the glottal phases and vocal quality. Consequently, we introduce another measure called the "abruptness index" to measure the information about the return phase of the glottal flow.

Table 2–4. Correlation coefficients across all utterances.

	B3	B2	B1	V3	V2	V1	M3	M2	M1
M1	0.7995	0.1530	0.7457	−0.5207	0.9687	0.5504	0.8607	0.9457	1.0000
M2	0.7224	0.2232	0.8203	−0.4217	0.9731	0.5609	0.9056	1.0000	
M3	0.6231	0.3244	0.8327	−0.1929	0.8891	0.2379	1.0000		
V1	0.4206	−0.1995	0.4858	−0.3752	0.5080	1.0000			
V2	0.7676	0.1488	0.7383	−0.4627	1.0000				
V3	−0.4187	−0.3896	−0.2588	1.0000					
B1	0.4877	0.3877	1.0000						
B2	−0.0001	1.0000							
B3	1.0000								

2.6.3.2 Abruptness index

The idea for this measure stems from the LF-model. In a study of the acoustic variability of the glottal source factors, Ahn (1991) concluded that the t_p, t_e of the LF model were the two most significant parameters correlating to vocal quality. Because t_e and t_c are the parameters controlling the rapidity of the return phase in the LF-model, they are accepted as an indicator of vocal abruptness. In Ahn's study, t_c was defined as the instant at which the amplitude of the modeled differentiated glottal flow dropped to 1% of its peak value.

Accordingly, t_c can be derived by solving the following equation

$$\frac{e^{-\xi(t_c - t_s)} - e^{-\xi(p_p - t_s)}}{1 - e^{-\xi(p_p - t_s)}} = 0.01 \tag{2–17}$$

where p_p is the pitch period and ξ can be obtained a priori by solving

$$\xi t_a = 1 - e^{-\xi(p_p - t_s)} \tag{2–18}$$

Since t_a and t_c form a mathematical mapping, we may say that the statistical significance of these two parameters stand on the same footing. Based on this understanding, we need to focus on only one parameter. It can be shown from Eq. (2–1) that

$$\frac{dE(t_e)}{dt} = \frac{E_e}{t_a} \quad or \quad t_a = E_e / \frac{dE(t_e)}{dt} \tag{2–19}$$

The equations above explicitly tell us that t_a can be readily obtained if we know the derivative of $E(t)$ at t_e. Since the instant t_e usually coincides with the largest value of $dE(t)$ and E_e is usually the minimum of $E(t)$, it will be convenient for us to calculate t_a using the following equation:

$$t_a = \frac{-\min(E(t))}{\max(dE(t))}\ \delta t \tag{2–20}$$

where δt denotes an infinitesimal time interval. For a discrete signal, this value, δt, can be substituted by the sampling interval, ΔT, provided that the interval is sufficiently small. If we define the vocal abruptness index, I_a, as the normalized t_a in percentage, then I_a becomes

$$I_a = \frac{-\min(\ E(n\Delta T)\ )}{\max(\ diff\ (E(n\Delta T))\ )}\ \frac{\Delta T}{P_P} \times 100\% \tag{2–21}$$

where ΔT is the sampling time, P_P is the pitch period, and *diff* stands for the difference function. It is obvious that I_a is readily obtained once $E(n\Delta T)$ is available.

The acquisition of the $E(n\Delta T)$ calls for an employment of the glottal inverse filtering technique, which is not always feasible. Thus, we demonstrate how to convert the residue signal into the differentiated glottal flow. As discussed in Section 2.3, a two-pole filter was employed to model the spectrum slope of the differentiated glottal flow $U_g'(n\Delta T)$. The transfer function of $U_g'(n\Delta T)$ is given as

$$U_g'(z) = \frac{e(z)}{1 - a_1 z^{-1} - a_2 z^{-2}} \qquad (2\text{–}22)$$

where $e(z)$ denotes the Z-transform of the residue. The glottal differentiated flow, $U_g'(n\Delta T)$, can be approximated by fitting the residue signal into the two-pole filter, of which the coefficient is derived by LP analysis of the speech signal. Substituting $U_g'(n\Delta T)$ for $E(n\Delta T)$ in Eq. (2–21), we can easily obtain the abruptness index (listed in Table 2–5 on page 61).

2.6.4 Vocal Noise

Much research has been directed to estimating vocal noise pertaining to a steady utterance or running speech. However, due to the limited capability of existing measures, the noise could only be presented in a form of signal-to-noise or harmonic-to-noise ratios that do not offer enough details of vocal noise. To gain a better understanding of vocal noise, we plan to examine the properties of noise from three aspects, namely, the signal-to-noise ratio (SNR), amplitude modulation, and noise spectra.

2.6.4.1 Noise extraction

In order to acquire the vocal noise, techniques for identifying and separating the prototype patterns (i.e., the standardized signal of one pitch period along an utterance) are required. The identification of pitch periods was accomplished by peak picking the DEGG signal. The separation of noise from the prototype can be achieved either in the frequency domain or in the time domain. There were two approaches that influenced us most in striving

to accomplish the noise extraction. One of the approaches was proposed by Yumoto et al (1982) who considered the noise as the deviation of a quasi-periodic speech signal. They first derived a prototype period of phonation by averaging the waveform of every period in a steady utterance. This prototype was then subtracted from the speech signal for each pitch period to yield the noise. The use of such an approach, however, requires that the subject's utterances have to be strictly steady for a number of periods. This may not be feasible in some cases. Thus, Kasuya et al (1986 a&b) proposed another measure by estimating the harmonic-to-noise ratio. The noise signal was isolated from periodical components either by using a comb filter or by collecting the non-harmonics in the spectrum of the analyzed signal. In such a method, any component not harmonically related to the fundamental frequency was classified as noise.

Although Kasuya's method was robust and efficient in computation, it was inappropriate from the practical point of view since it took account of the jitter and shimmer. In our proposed speech production model the vocal noise is considered to be an independent module. This design concept requires that the uncorrelated factors, such as jitter and shimmer, has to be segregated from the real noise source. Thus, we adopt Yumoto's idea in a modified form. First, we standardize the power and length of every pitch period of the integrated residue using the method discussed in Section 2.5.4 before acquiring the prototype. Then we estimate the vocal noise by minimizing the least square difference, $E(S_i, S_p)$, between the prototype signal S_p and the analyzing signal S_i.

$$E(S_i^l, S_p) = \sum_{k=0}^{N-1} [C^l(S_i(k)) - \gamma S_p(k)]^2 \qquad (2\text{–}23)$$

where C^l denotes the circulation shift with an l lag, and N is the length of the standardized pitch period. The purpose of using C^l is to rectify the phase discongruity between S_i and S_p. The scale factor γ is then determined by setting

$$\partial E(S_p^i, S_p)/\partial \gamma = 0, \tag{2–24}$$

which leads to

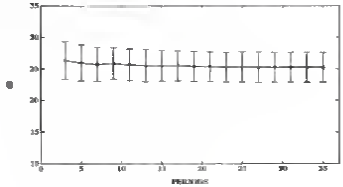
$$\gamma = \frac{\sum_{k=0}^{N-1} C^m(S_i(k)) S_p(k)}{\left(\sum_{k=0}^{N-1} S_i^2(k) \sum_{k=0}^{N-1} S_p^2(k) \right)^{1/2}} \tag{2–25}$$

The lag m that yields the maximum γ is chosen be the correct phase offset. Thus, the noise signal becomes

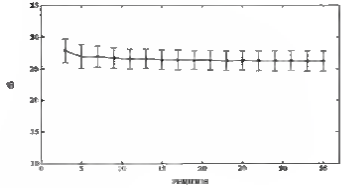
$$n(k) = C^m(S_i(k)) - \gamma S_p(k) \tag{2–26}$$

A problem of this approach above is related to determining an appropriate number of periods to form an analysis window. The prototype derived from a short window is statistically unreliable. On the other hand, it is unlikely that a steady phonation is maintained throughout the utterance. Therefore, we have to examine the influence of the number of periods with regard to the noise measure. An empirical but sensible criterion for selecting the analysis window is to search for the minimum period that gives a small standard deviation. Here we use three different utterances as a pilot experiment. As shown in Figure 2–14, the standard deviations of three samples are relatively large when the analysis window is small. When the analysis window is increased to more than 15 periods, both the standard deviations and mean values become stable. Thus, the prototype is calculated using a window of 15 consecutive periods, in which the current period is located at the center of the window. The selected number is somewhat smaller than that suggested by other researchers (Yumoto et al., 1982, Titze et al., 1987, Eskenazi et al , 1990). We reason that this result is due to the resolution enhancement and pitch standardization.

(a) modal (M1)



(b) vocal fry (V1)



(c) breathy (B1)

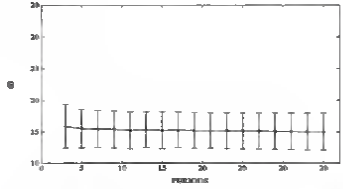


Figure 2–14 Variation of SNR versus number of analyzed periods (range of each error bar = [–std, std]).

Another problem causing concern is the fluctuation of low frequency power. Because the air escaping from the lungs is not continual, a frequent change of low frequency components is anticipated. Fortunately owing to the fact that the noise in the low-frequency region is perceptually masked by the harmonics of the fundamental frequency (Childers and Lee, 1991), we can apply a notch filter to eliminate the low frequency components without disturbing the perceived quality. The cut-off frequency of this highpass zero-phase filter can be designed to adapt to the current pitch period P_i such that low-frequency components below 500 Hz are sufficiently suppressed and high frequency components are not affected. The frequency response of the highpass filter is given by

$$H(z) = \frac{2-\alpha}{2} \frac{1-z^{-1}}{1-(1-\alpha)z^{-1}} \tag{2-27}$$

where $\alpha = \frac{0.2P_i}{512}$ provides the adaptation for the ith pitch period of length P_i. The number, 512, is the length of the pitch period after interpolation. The scalar factor $(2-\alpha)/2$, in Eq. (2–27) is to make the magnitude unity at the one-half the sampling frequency. Eventually, the desired noise signal is the result after passing $n(k)$ through the notch filter.

2.6.4.2 Properties of vocal noise

Once we get the desired noise, the properties to be evaluated are:

- Signal-to-Noise ratio

The Signal-to-Noise Ratio (SNR) for the ith pitch period is calculated as

$$SNR_i = 10\log_{10}\left[\frac{\gamma^2 \sum_{k=1}^{N} S_i^2(k)}{\sum_{k=1}^{N} n_i^2(k)}\right], \tag{2-28}$$

and the SNR's for the nine utterances are listed in Table 2–5

Table 2–5 Mean values and standard deviations ($\pm$ std) of the acoustic measures for each subject.

	a_1	a_2	I_0	SNR [dB]
M1	−0.561 ± 0.043	−0.175 ± 0.016	0.890 ± 0.165	25.466 ± 2.450
M2	−0.780 ± 0.037	−0.026 ± 0.043	1.290 ± 0.169	23.184 ± 2.138
M3	−0.757 ± 0.454	−0.139 ± 0.026	1.488 ± 0.255	25.329 ± 1.793
V1	−0.646 ± 0.051	−0.079 ± 0.021	0.681 ± 0.137	26.468 ± 1.538
V2	−0.584 ± 0.099	−0.032 ± 0.063	0.858 ± 0.291	24.822 ± 2.543
V3	−0.235 ± 0.151	0.177 ± 0.097	0.263 ± 0.090	14.998 ± 3.401
B1	−0.933 ± 0.090	−0.036 ± 0.081	1.371 ± 0.288	15.282 ± 2.945
B2	−1.263 ± 0.150	0.272 ± 0.148	2.645 ± 1.167	6.209 ± 2.536
B3	−1.657 ± 0.093	0.682 ± 0.091	5.135 ± 1.224	14.964 ± 2.703

- Amplitude modulation

As shown in Figure 2–15, the amplitude modulations is obtained by averaging the magnitude of the noise signal over all periods.

- Noise spectrum

The spectrum of the noise signal is computed using the FFT. Though the length of every pitch period has been expanded to 512 samples, the frequency resolution of the FFT sequence still depends on the actual fundamental frequency. Due to the fact that the fundamental frequency may change from period to period, the resolution of each FFT sequence is therefore different from each other. Thus, we apply the biharmonic interpolation on the FFT sequences to achieve a unique frequency resolution. The individual FFT spectra are then averaged to yield an estimation of the noise spectrum. Figure 2–16 shows the period spectra for different voice types.

2.6.4.3 Brief summary

To summarize, we present the noise extraction algorithm as a flowchart in Figure 2–17. We recall that the noise is extracted from the residue signal, which is obtained using techniques addressed in previous sections. The overall procedure is tedious, but is straightforward and easy to implement.

2.7 Discussion

The results we have gained so far can be summarized in four aspects.

(1) It was found that the perturbation noise can be modeled by a zero-mean Gaussian process with a low-frequency drift. Measures that sufficiently deemphasized the drift could be used to characterize the source perturbations. In particular, we have used the %jitter and %shimmer to indicate the standard deviations of the noise sources. The results of measured perturbations with respect to three voice types, in general, were consistent with other researchers' reports, i.e., vocal fry and breathy voices exhibit higher perturbations. We also

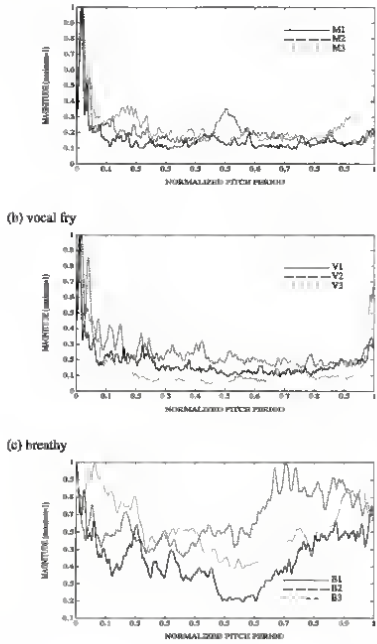
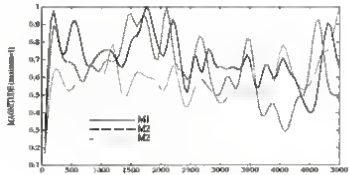
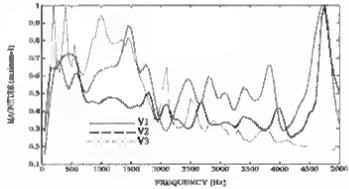


Figure 2–15 Amplitude modulations of vocal noise for different voice types.

(a) modal



(b) vocal fry



(c) breathy

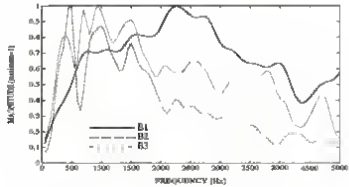


Figure 2–16. Spectra of vocal noise for different voice types.

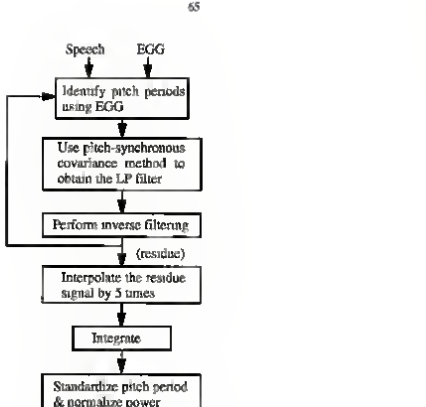


Figure 2–17. Schematic flowchart for noise extraction.

found that the smaller perturbations in vocal fry and breathy voices corresponded to low pitch subjects. Interpreted from a psychoacoustic perspective, such values would have different impacts to the perception of vocal quality (Wendahl, 1963). Furthermore, loudness, a perceptual descriptor of the intensity, was reported to be a nonlinear function for various frequencies (Robinson and Dadson, 1956). These factors confound the study of voice quality merely on the basis of quantitative measures. To achieve a thorough understanding of vocal quality, the research scope should cover speech perception as well (Flanagan, 1972a; Bladon and Lindblom, 1981; Hermansky et al., 1985, Wang et al., 1991).

(2) A comparison of the spectral tilt of the source can be performed by visually inspecting the frequency responses of the two-pole filter model. As shown in Figure 2–18, the spectral tilt is moderate, relative flat, and steep for vocal fry, modal and breathy voices, respectively. A simple quantitative measure can be achieved by comparing the first coefficient a_1, since the coefficient a_1 and the poles $z_{1,2}$ of the modeled filter have the following relation:

$$\begin{aligned} a_1 &= -2|z_1|\cos\theta \approx -2|z_1| && \text{if } \theta \neq 0, \pi, \\ &= -|z_1 + z_2| && \text{if } \theta = 0, \pi \end{aligned} \qquad (2\text{–}29)$$

where $\theta = \tan^{-1}\left[\frac{\Im(z_1)}{\Re(z_1)}\right]$

Hence, the value of a_1 can be used to indicate how close the poles are to the unit circle. A larger a_1 corresponds to a flatter spectral tilt and broader bandwidth. According to the data in Table 2–5, the values of a_1 for different voice types exhibited the following inequality:

$$\{|a_1|_{\text{vocal fry}}\} > \{|a_1|_{\text{modal}}\} > \{|a_1|_{\text{breathy}}\}$$

This result is congruent with the previous observation in Figure 2–18 and with the conclusions shown in Table 2–2.

(3) As the relation between the residue signal and glottal flow was unveiled, the glottal phase characteristics could be traced back from the residue signal. We have studied

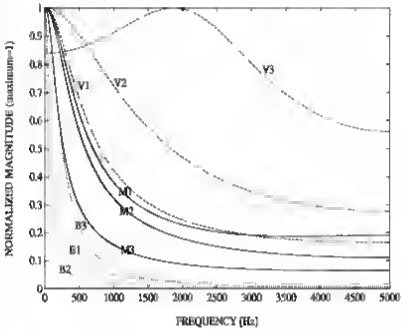


Figure 2–18. Frequency responses of two-pole filters for nine subjects

the similarities and differences among the nine integrated residue signals by examining their correlation coefficients. It was found that no similarity exists except for modal voices. This result suggests that the glottal phase characteristics cannot be described with a general pattern. On the contrary, the abruptness index showed an advantage for characterizing voice types. The values of the abruptness indexes for three voice types are presented in descending order as vocal fry, modal, and breathy voices. Such a measure may enable researchers to classify voice types with considerable convenience. It is worth noting that the meaning of this measure can be interpreted from various aspects. In the time domain, it is related to the temporal transition from the maximum glottal closure instant to the glottal opening. In the frequency domain, it indicates the spectral slope of the glottal source. From the point of view of the residue signal, it corresponds to the peak factor of the main excitation pulse.

(4) The estimated SNR's for various voice types supported the earlier finding of other researchers that breathy voices, in general, were accompanied by the largest vocal noise. Such a noise level was distinctive enough to underscore its role in source modeling. An intriguing observation with the breathy voices is that the standard deviations of corresponding SNR's are as small as that of modal voices. This result suggests that a steady noise source would be appropriate to model the vocal noise for modal and breathy voices.

As displayed in Figure 2–14, the noise spectra for different phonations were fairly flat, suggesting that the noise for the integrated residue is white. However, for the purpose of speech synthesis, the noise source has to be pre-emphasized by a highpass filter before applying to a LP synthesizer.

The amplitude modulation of the vocal noise generally resembles the magnitude of the integrated residue (Figure 2–15). However, the high amplitude modulation near the glottal closure may also be ascribed to the phase misalignment. Notice that there are two types of randomness presented in the residue signal: one is the epoch variation caused by the vocal fold closure, and the other is the variation due to the airflow turbulency from the lungs (Kang and Everett, 1985). The integration with respect to the residue makes the phase

adjustment procedure in favour of the airflow, thus increasing the degree of mismatching for the epoch variation. The accusation above can be further verified by the subsequent derivation. Suppose there is a phase offset, θ, between two signals, $s(t)$ and $s_1(t)=\Re\{s(t)e^{-j\theta}\}$, then the difference $e(t)$ is

$$\begin{aligned} e(t) &= s(t) - s_1(t) \\ &= s(t) - \Re\left\{s(t)e^{-j\theta}\right\} \\ &= 2s(t)\sin^2(\frac{\theta}{2}) \\ &= s(t)\frac{\theta^2}{2} \qquad \text{, for } |\theta| \ll \frac{\pi}{2} \end{aligned} \tag{2–30}$$

Clearly from Eq (2–30), the error $e(t)$ is proportional to the signal, resulting in the similarity between the amplitude modulation of the extracted noise and the magnitude of the integrated residue.

So far it is undetermined whether the amplitude modulation is an artifact of the analysis method or is a primitive feature of the glottal source. We will revisit this issue in Chapter 4. But one thing is certain here, that is, the quality of synthetic speech is affected by the simulation of vocal noise.

2.8 Conclusion

In this chapter we have explored the acoustical features within the source-filter theory. The properties of the glottal source were primarily extracted from the integrated residue signal, which was obtained by making use of the pitch synchronous LP analysis with the aid of the DEGG signal. We demonstrated the analysis methods using sustained vowels, */a/*'s, of three voice types, i.e., modal, vocal fry, and breathy voices. The roles of many existing acoustic measures were carefully investigated. Although more extensive investigations are needed in order to establish statistical significances of model parameters, the results of our study provided a basic understanding of source variations as well as their manifestations in the acoustic measures. More important, the capabilities of extracting the

glottal source properties using LP analysis were substantiated. The competence of LP method in speech analysis suggests that a high quality LPC synthesizer is achievable. Of course, this is under the assumption that the properties of the glottal source are faithfully preserved during the analysis and synthesis. To achieve such a requirement, two important features that are usually ignored in many LP synthesizers, i.e. the vocal noise and the glottal phase characteristics, have to be incorporated into the source model.

CHAPTER 3
SOURCE MODELING

Since it was introduced in the late 1960s, the linear predictive coding (LPC) technique has been extensively used in speech processing and coding (Rabiner and Schafer, 1978). Speech synthesizers considered in the class of LPC coders use a slowly time-varying all-pole filter to model the composite spectral characteristics of the glottal flow, vocal tract and lip radiation. The excitation for this all-pole filter is a spectrally flat signal with quasi-periodic phases for voiced speech and random phases for unvoiced speech. In this study, we apply a sixth order polynomial model to delineate the phase characteristics of voiced source excitation. Source features extracted by this model are further compressed through a vector quantization technique. A 32-entry glottal codebook is derived by quantizing the voiced samples uttered by 20 subjects. On the other hand, a 256-entry stochastic codebook is generated for unvoiced speech synthesis. However, unlike the glottal codebook, codewords in the stochastic codebook are simply taken from a Gaussian noise source.

3.1 Review of Previous Research

Over the years, various types of excitation have been proposed to drive the synthesis filter to produce speech. In the conventional pitch-excited LPC vocoder (Atal and Hanauer, 1971), the excitation signal is either an impulse train for voiced speech or a random noise for unvoiced speech. The quality of synthesized speech in some applications is judged as unnatural due to incorrect voicing decisions, poor spectral resolution and oversimplified excitation functions (Wong, 1980, Kahn and Garst, 1983).

The use of rather sophisticated excitation functions such as the Multi-Pulse (MP), Code-Excited (CE) or their relatives (Atal and Remde 1982, Schroeder and Atal, 1985, Singhal and Atal, 1989, Rose and Barnwell, 1990) can result in high-quality synthetic speech if the synthetic excitation is described sufficiently well by adequate number of codewords or pulses. Coders using this type of excitation go beyond spectral analysis and pitch estimation. Features not representable by predictive filters can be recovered by formulating the excitation signal. That is, the excitation signal is formed by searching for the best candidate in a given set of innovative sequences by minimizing the spectrally weighted difference between the original and the synthesized speech signals.

In fact, the ideal excitation for LP synthesizers is the residue signal obtained by inverse filtering of the original speech signal. Attempts have been made to encode and transmit the residue signal in many coding systems (Un and Magill, 1975, Dankberg and Wong, 1979). But little research effort has been directed to extracting the features of the residue signal. In 1978 Wong and Markel constructed a prototype excitation pulse by inverse filtering the differentiated glottal flow of the vowel */a/*. Although this excitation pulse was intentionally designed to reduce the buzziness of synthesized speech, both quality and naturalness, as expected, were improved due to the preserved glottal characteristics. However, the excitation pulse presented in their experiment has certain drawbacks. First, it is feasible only when the fundamental frequency is below 160 Hz. Second, a single prototype excitation pulse is not likely to suit all the situations since glottal features for various speakers and phonations can vary considerably.

The importance of glottal characteristics for speech synthesis was also demonstrated by Bergstrom and Hedelin (1989). Finding the similarity between the residue and the second derivative of the glottal pulse, they incorporated the glottal pulse into a CELP coder by adding an extra codebook. The resultant quality of synthetic speech was reported to be favored over the quality produced by the primitive CELP coder. Recently, the incorporation of the residual features by means of excitation codebooks also gained a certain degree of

successfulness in synthesizing natural speech at 2.4 Kb/s (Haagen et al. 1992, Zhang and Chen, 1992).

Other attempts to replace the residue by stylized pulses appear in the work by Sambur et al. (1978) and by Childers and Wu (1990). Among the tested pulses, the differentiated electroglottograph (DEGG) signal was found to produce good quality. Such a result occurred because the DEGG signal reflects the glottal characteristics and has a rather flat spectrum.

In contrast to the foregoing approaches, many researchers have adopted a "divide-and-conquer" strategy to depict the residue. Some divided the spectrum of the residue into several frequency bands and examined the corresponding spectral characteristics for each band (Makhoul et al, 1978, Kwon and Goldberg, 1984, Griffin and Lim, 1988, McCree and Barnwell, 1991). The excitation signal was then formed by summing the subband components under the constraint that the resulting excitation must exhibit a flat spectrum. If there were only two spectrum bands to be specified, the model was often referred to as the mixed excitation since it resulted in a mixture of low frequency pulses and high frequency noise. If the number of divided bands matched that of pitch harmonics, this type of excitation became a superposition of sinusoids and was named under its general properties as either harmonics or sinusoids (Trancoso et al., 1990). On the other hand, such a "divide-and-conquer" strategy was also considered by researchers for use in the time domain. Sreenivas (1988) parsed the residue signal into three parts, i.e., high energy pulses, a low energy smooth component, and a random noise component. Each component was acquired by using a distinctive feature. For instance, the high energy pulses were found based on a error minimization scheme similar to MPLP coders. After subtracting the pulses from the residue, the smooth component was calculated by vector quantization. Likewise, the noise component was determined by codeword searching as in CELP coders. Such an approach has proven useful for speech coding in the range of 9.6 Kb/s. Sukkar et al. (1989) decomposed the residue into a set of orthogonal functions called Zinc functions. They

claimed that the Zinc-function is superior to the Fourier expansion for modeling the residue in the mean square error sense. However, even though both the frequency and time domain approaches offered better synthetic quality, none of the above mentioned models provided a clue to describe the glottal features parametrically.

From the above discussion, it appears that the quality of synthesized speech can be improved once we attend to the basic features of the residue signal. Our investigation showed that the residue was closely related to the glottal volume velocity via the glottal shaping filter (see Section 2.3). In fact, Kang and Everett (1985) have demonstrated how to improve the quality of the pitch-excited LPC vocoder through the exploitation of the amplitude and phase spectra of the residue. It was also reported that high-quality LP synthesis could be achieved by introducing an extended filter which captured some of the glottal phase characteristics (Caspers and Atal, 1987; Hedelin, 1988). The improvement due to the appropriate modeling of glottal source is more evident when a glottal flow model is applied to the formant synthesizers (Rosenberg, 1971; Holmes, 1973; Klatt, 1980; Pinto, et al. 1989), but such perceptually important features have not been widely considered in LP synthesizers. Our primary goal is to design an efficient excitation model to simulate the residue so that we may achieve high-quality natural-sounding speech production using such an excitation model.

3.2 Excitation Source

In a manner similar to that adopted in the traditional LP synthesizer, we classify the excitation function into two categories, i.e., voiced and unvoiced. Accordingly, two different strategies are employed to analyze and process the speech signal.

3.2.1 Voiced Segments: Excitation Pulse

In Section 2.3, we have shown that the phase characteristics of a glottal flow waveform could be retrieved from the residue signal. However, since the zero-reference

level of the glottal flow has been destroyed due to the inverse filtering and integration, source models that specify the differentiated glottal flow are not suitable for modeling the integrated residue. We therefore propose a new model to code the integral of the residue. This model is described by a sixth order polynomial $f(x) = \sum_{i=0}^{6} c_i x^i$, which is specified within the interval [0,1] subject to three constraints listed below

1. $f(0) = -1$ (3–1)

2. $f(1) = f(0)$ (3–2)

3. $\int_0^1 f(x)dx = 0.$ (3–3)

where the interval boundaries, 0 and 1, correspond to the glottal closure instants (GCI). The order of the polynomial is empirically chosen to be six because it sufficiently describes the integrated residue without causing rank deficiency.

The purpose of the constraints is as follows. The first constraint is used to normalize the magnitude of the largest negative peak. The second constraint is to ensure the circular continuity between consecutive periods. It is also equivalent to the following expression

$$\int_0^1 f'(x)dx = 0, \tag{3–4}$$

which indicates that the d.c. component in the residue signal is eliminated. The third constraint is established to avoid any low-frequency modulation.

Because of these constraints, only four degrees of freedom are available in the polynomial even though seven coefficients exist. To acquire the polynomial coefficients under such constraints, we can introduce Lagrange multipliers and solve a set of equations as in an optimal control system. Nonetheless, the main purpose of these constraints is not

to limit the dynamics of the polynomial coefficients while carrying out the optimization. Instead, the constraints are just used to regulate the polynomial waveform. They can also be satisfied by adjusting a tentative polynomial, which is calculated based on a least square fit. Here we apply a weighting function to emphasize the polynomial fitness around the GCI since this region is directly related to the primary excitation pulse. The weighting function is given by

$$W(x) = \begin{cases} 200x^2 - 40x + 3 & \text{for } 0 \le x \le .1 \quad [1] \\ 1 & \text{for } .1 < x < .8 \quad [2] \\ 25x^2 - 40x + 17 & \text{for } .8 \le x \le 1 \quad [3] \end{cases} \tag{3–5}$$

and is displayed in Figure 3–1. In practice, the weighting function can also reduce the chance of rank deficiency while we perform the polynomial fit.

Once we obtain the tentative polynomial, the first constraint can be achieved by normalizing all the coefficients with respect to C_0, i.e.,

$$C_i = -\frac{C_i}{C_0} \qquad \text{for } i = 0, 1, 2, 3, 4, 5, 6 \tag{3–6}$$

The second constraint can be satisfied by seeking a value v close to 1 such that $f(v)$ is 1. Accordingly, the polynomial coefficients are revised as

$$C_i = C_i v^i \qquad \text{for } i = 1, 2, 3, 4, 5, 6. \tag{3–7}$$

The solution regarding the third constraint turns out to be a procedure for removing the d.c. level. We can modify the constant C_0 to accomplish this requirement

$$C_0 = -\sum_{i=1}^{6} \frac{C_i}{i+1} \tag{3–8}$$

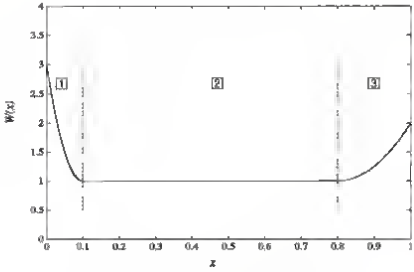


Figure 3–1 Plot of the weighting function $W(x)$

Thus, the resultant integral for a period becomes

$$\int_0^1 f(x)dx = \sum_{i=0}^{6} \frac{C_i}{i+1} = 0 \qquad (3\text{-}9)$$

It is important to note that when applied to the tentative polynomial the above-mentioned adjustments shall be arranged as Eq (3-7), then Eqs (3-8) and (3-6) in order to prevent any further conflict among the constraints.

3.2.1.1 Vector quantization

Like many other glottal source models, the polynomial model only provides a rough description of the glottal phase characteristics. The lack of detailing of the glottal phase may lead to a degradation in quality of synthetic speech. However, in a study concerning the influence of glottal flow waveforms on the quality of voiced synthetic speech, Ronsenberg (1971) concluded that only gross source features are required to preserve the quality whereas temporal and spectral details are less important. Assertions regarding the phase characteristics were further supported by other researchers (Atal and David, 1979, Hedelin, 1986). Their results lead us to speculate that the glottal excitation acquired by our model may provide sufficient discriminatory information in order to synthesize good quality speech. It is noted that vector quantization techniques have demonstrated good performance in compressing LP features with a relatively low bit rate (Linde et al., 1981; Gray, 1984). We believe that the glottal phase characteristics portrayed by our source model could be more concise via an appropriate vector quantizer, at least in terms of perceptual quality.

In a general sense, the quantization is a process for converting a continuous-amplitude sample into one of a set of discrete-amplitude samples suitable for storage and communication in a digital system. The process is known as scalar quantization, if each individual sample is quantized independently. When a block of samples, usually defined as a vector, is quantized jointly, the process is termed vector quantization.

Given a K-dimensional Euclidean space R^k, a vector quantizer considered a criterion partitioning R^k into a finite subset Y of R^k, where $Y = \{y_i : i=1,2, ..., N\}$ is the set of reproduction vectors and N the number of vectors in Y. The set Y is called a codebook and its elements are called codewords or codevectors. In principle, the codeword y_i is chosen to minimize the average distortion for each quantized cell. The distance between any input vector and its corresponding codeword is known as the distortion. Once these codewords are established, any input vector is then assigned to a particular codeword based on minimum distortion for optimal representation. More specifically, pattern vector x is encoded by codeword y_i if the distance between those two vectors is less than the distance to any other codeword, i.e.,

$$d(x, y_i) < d(x, y_j), \qquad j \neq i \; ; i,j = 1 \; ... \; N \qquad (3\text{–}10)$$

where the function d denotes the distance measure, and N is the number of codewords. A major advantage with the vector quantizer is that it often reduces the number of bits required to represent the input vector under a specific distortion measure. Indeed, this advantage can be formally proven through mathematical derivations. According to the Shannon rate-distortion theory, the vector quantizer always achieves higher data compression ratios than any coding scheme based on the scalar quantities for a given transmission bit rate. Because of this, during the past decade, the vector quantization has received much attention as a data compression technique for encoding data in information intensive fields such as image and speech signals.

A vital step in establishing the vector quantizer is generation of an accurate codebook. Here the word "accurate" stands for having minimum distortion. The accomplishment of this step requires a criterion to quantify the Euclidean space and a distortion measure to define the performance of a quantizer. There are two distortion criteria commonly adopted for vector quantizers, namely, either minimizing the average quantization error or maximizing the codebook entropy defined as

$$E = -\sum_{i} P_i \log_2(P_i) \qquad (3\text{–}11)$$

where P_i is the relative frequency with which codeword i is used to encode the sample vectors. While it may seem intuitive to demand a quantizer to minimize the average distortion, the most efficient way to quantize the vector space is to let each quantized cell (also known as "cluster" in some literature) consist of the same entropy. Conceptually, minimizing the average quantization error can be viewed as a scheme performing a geometric division of the vector space, while maximizing the entropy is a scheme to achieve a popular division of the vector space. Our philosophy of quantizing the vector space is to minimize the quantization error but at the same time to maximize the selected frequency of each codeword. It appears that the sum of intra-cluster distortion serves as a proper criterion for cluster splitting because this criterion takes both geometric and popular division properties into account (Tou and Gonzalez, 1974; Nyeck and Tosser Roussey, 1992). Under such a criterion, clusters containing excessive training sample vectors are more likely to be split despite their intra-cluster distortions are low. Hence the codebook space is not wasted in accommodating unusual pattern vectors of glottal phase signals.

A perfect partition for the pattern space may be quite difficult to accomplish, although it is theoretically obtainable when the distortion measure is specified and the probability density function of input vectors is known. Such a difficulty, however, can be circumvented by making use of long training sequences that approximately represent the probability density function. Thus, if the vector process is ergotic and stationary, averaging the distortion for a large amount of training vectors is equivalent to applying the probabilistic model to the underlying process. Since each vector is mapped into only one particular codeword, the codewords themselves may be established through clustering techniques. In fact, the optimal codeword is just the centroid of its associated clusters subject to a selected distortion measure. This implies that the cluster analysis algorithms in pattern recognition literature, such as *K*-means, ISODATA, DYNOC, and some neural-net techniques can be

used to categorize the training vectors into clusters or, equivalently, to determine the hyperplane partitioning the clusters (Tou and Gonzalez, 1974, Tou, 1979, Pao, 1989).

3.2.1.2 Maximum decent algorithm

In this study we generate a 32-entry codebook using a maximum decent algorithm (Ma and Chan, 1991). We note that the size of the codebook is just tentative. This number is, in general, determined by the transmission system and the desired compression ratio.

The maximum decent rule says that the clusters are chosen one at a time attempting to achieve a maximum reduction of the sum of the distortions. As illustrated in Figure 3–2, we begin the splitting routine by placing all vectors in a global cluster. After forming the first two clusters, we compare the reduction functions, R_1 and R_2, of the two new clusters and then split the one giving the larger reduction. To generalize the preceding procedures, let us consider the case of forming $n+1$ clusters based on a set of n clusters. The cluster S_m ($m \leq n$) is split into two new clusters if R_m is the largest among all the R_i's of the n clusters. Hence the set of $n+1$ clusters is the one that gives the maximum decent distortion when formed from the set of n clusters. The algorithm iterates until the desired number of clusters is obtained. Finally, the centroids of the clusters are taken as the codewords.

The advantages of using the maximum decent algorithm include: (1) computation time is significantly reduced since only the R_i's of the two newly formed clusters need to be computed while all other clusters have been calculated in the previous iteration; and (2) empty clusters are prevented since it is impossible for a single-member cluster to be chosen for splitting.

3.2.1.3 Cluster splitting

Since each codeword represents the centroid of a specific cluster, the size of the codebook equals the number of clusters partitioned in the pattern space. We adopt a splitting technique to carry out the cluster partition. This technique, in general, is not guaranteed to

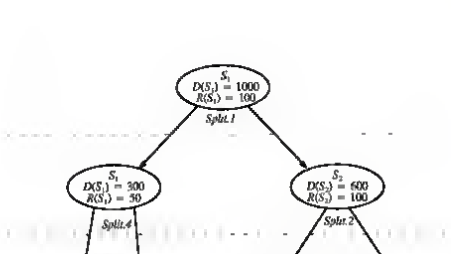


Figure 3–2. Cluster splitting using the Maximum Decent method.
$D(S_i)$=sum of distortions,
$R(S_i)$=reduction of distortion due to cluster splitting.

provide an optimal solution, but it gives satisfactory results even with a binary search coding scheme (Buzo et al., 1980). Steps for splitting each given cluster are summarized as follows.

Step 5. Assign the initial centroids by using the extreme-point approach, which will be discussed in Section 3.2.1.3.1.

Step 6. Partition the cluster vector on the basis of minimum distortion, i.e.

$$\text{if } d(x_j, y_1) < d(x_j, y_2), \quad x_j \in S_{l1},$$
$$\text{otherwise,} \quad x_j \in S_{l2}$$

Step 7. Obtain the new centroids by

$$y_1^l = \frac{1}{N_1} \sum_{x_i \in S_1} x_i,$$

$$y_2^l = \frac{1}{N_2} \sum_{x_i \in S_2} x_i$$

where N_1 and N_2 are the number of vectors assigned to S_1 and S_2, respectively. The superscript l denotes the number of iterations.

Step 8. Calculate the reduction of distortion R_l^l due to splitting as

$$R(S_l) = D(S_l) - [D(S_{l1}) + D(S_{l2})],$$

$$\text{If } \frac{R_l^l - R_l^{l-1}}{R_l^l} > 10^{-5}$$
then Go to Step 2,
else Terminate

The outcome of the cluster analysis, in general, will be affected by three factors, namely the initial centroids, distortion measure, and the geometric properties of the training vectors. The geometric properties reflect the distribution of feature patterns and can be adjusted by properly selecting the training vectors. At this stage, we can assume that the

selected training vectors are exemplary both in completeness and equilibrium. Thus, we are concerned only with the initial centroids and distortion measure.

3.2.1.3.1 Initialization of centroid

Several methods for determining the initial codewords exist. We may simply choose the first two training vectors as our initial centroids, similar to the manner used in the *K*-means method. However, simply choosing the first two vectors will not produce an accurate result if these two vectors are close to each other. Intuitively, one would like these two vectors to be well separated. We therefore, assign the two initial centroids using the following approach. Let $\{x_1, x_2, x_3, ..., x_N\}$ be the N sample vectors. The mean vector is given by

$$z_0 = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{3–12}$$

Using z_0 as a reference vector, we first find a vector x_m that is farthest from z_0. That is,

$$d(x_m, z_0) > d(x_i, z_0), \qquad \text{for } i \neq m, \ \ i, m = 1, \ \ , N \tag{3–13}$$

This vector x_m is selected as one of the extreme vectors. The other is determined by searching for the vector that is farthest from x_m.

3.2.1.3.2 Distortion measure

As mentioned earlier, the feature space consists of the polynomial coefficients. We adopted the Euclidean distance as the distortion measure, which is defined as

$$d_{ij} = \int_0^1 (f_i(x) - f_j(x))^2 dx \tag{3–14}$$

where d_{ij} is the resulting distortion of two arbitrary polynomials, $f_i(x)$ and $f_j(x)$. The centroid, $f_k(x)$, of a cluster, S_k, is chosen as

$$f_k(x) = \frac{1}{N_k} \sum_{f_i \in S_k} f_i(x) \tag{3–15}$$

where N_k is the number of vectors inside S_k. Thus, the sum of distortions $D(S_k)$ for the cluster S_k is given by

$$D(S_k) = \sum_{f_i \in S_k} \int_0^1 (f_i(x) - f_k(x))^2 dx. \tag{3–16}$$

Let P_c define the vector of the polynomial coefficients, $(f_i(x) - f_k(x))$, in a descending order. The polynomial multiplication of $(f_i(x) - f_k(x))^2$ is equivalent to convolving P_c with itself, i.e., $P_{ac} = P_c * P_c$, where [*] denotes the convolution operator and P_{ac} is the coefficient sequence of resulting polynomial. After solving the integral function, Eq. 3–16 becomes

$$D(S_k) = \sum_{f_i \in S_k} \sum_{k=1}^{n} \frac{P_{ac}(k)}{(n+1-k)} \tag{3–17}$$

where n is the number of coefficient of P_{ac}. In our case, $n = 13$.

3.2.1.4 Codebook training

In order to reflect the source variation caused by factors such as stress and intonation, we use sentences instead of sustained vowels for training the glottal codebook. The selected sentences are: (1) "We were away a year ago." spoken by 16 subjects, and (2) "Early one morning a man and a woman ambled along a one mile lane," spoken by 4 subjects. In both cases, the numbers of both male and female subjects are equal. The data base is shown in Table 3–1. The resulting codebook is given in Table 3–2. The inclusion of nasals, as in the second sentence, is intended to compensate for the deficiency of the all-pole model by attributing zero (anti-formant) characteristics to the source model. Although the set of training samples does not consist of all possible voiced sounds, source properties are still considered representative since the supraglottal loading effects are removed by the inverse filter and source characteristics are presumably the only remaining ingredients.

Table 3–1 Data base for codebook training.

Initials	Sex	# of pitch periods	sentence	Initials	Sex	# of pitch periods	sentence
STR	M	400	(1)	CLW	F	200	(1)
JJS	M	318	(1)	AHZ	F	168	(1)
JMS	M	294	(1)	C2G	F	227	(1)
JTO	M	122	(1)	P2B	F	258	(1)
DMH	M	155	(1)	JEW	F	277	(1)
ESC	M	115	(1)	JLH	F	278	(1)
DCD	M	282	(1)	MBK	F	286	(1)
TLB	M	275	(1)	PXS	F	136	(1)
DRW	M	383	(2)	CAP	F	591	(2)
MJS	M	337	(2)	LAD	F	761	(2)

3.2.2 Unvoiced/Silence Segments: White Noise

For simplicity, we treat silence as unvoiced speech since the power level of the silence segments is so low that any modeled errors can be attributed to background noise. Similar to the idea adopted in voiced excitation, a stochastic codebook is used as the excitation source for unvoiced speech. This implies that the residue is simulated using a finite number of innovation sequences subject to a given fidelity criterion. The use of such innovation sequences is motivated by the CELP coders, of which the stochastic codebook has been known to produce better unvoiced speech than voiced speech for low bit coding (Schaltherb and Lacroix,1989). But in contrast to the fundamental structure of the CELP coder, the commonly used long-term predictor is dropped here since pitch harmonics are unnecessary in unvoiced speech.

Basically, the size of the codebook is determined by three factors, namely, the transmission rate, computational complexity and frame update rate. Due to the lack of an

Table 3–2. Content of glottal codebook.

Codeword	C_6	C_5	C_4	C_3	C_2	C_1	C_0
							$\times 10^3$
1	−1.1102	3.3856	−3.9522	2.2104	−0.5999	0.0663	−0.0010
2	−0.7313	2.4157	−3.0987	1.9110	−0.5592	0.0625	−0.0010
3	−0.0596	0.2469	−0.3923	0.2859	−0.0949	0.0140	−0.0010
4	−0.1922	0.6130	−0.7384	0.4189	−0.1193	0.0181	−0.0010
5	−0.4829	1.5736	−1.9304	1.0931	−0.2836	0.0301	−0.0010
6	−0.0786	0.3093	−0.5154	0.4201	−0.1586	0.0233	−0.0010
7	−0.6844	2.1580	−2.5733	1.4361	−0.3778	0.0413	−0.0010
8	−0.3847	1.2809	−1.6114	0.9343	−0.2446	0.0254	−0.0010
9	−0.2608	0.8897	−1.1955	0.7797	−0.2443	0.0312	−0.0010
10	−0.4336	1.3968	−1.7093	0.9935	−0.2798	0.0344	−0.0010
11	−0.2545	0.8522	−1.0869	0.6479	−0.1806	0.0219	−0.0010
12	−0.0643	0.2436	−0.3718	0.2830	−0.1105	0.0201	−0.0010
13	−0.4961	1.5986	−1.9584	1.1270	−0.3056	0.0345	−0.0010
14	−0.0257	0.1359	−0.2610	0.2250	−0.0905	0.0163	−0.0010
15	−0.7603	2.4966	−3.1414	1.8677	−0.5172	0.0545	−0.0010
16	−0.5183	1.7980	−2.4151	1.5421	−0.4574	0.0507	−0.0010
17	−0.9932	3.0734	−3.6201	2.0159	−0.5322	0.0562	−0.0010
18	−0.3318	1.0752	−1.3293	0.7789	−0.2212	0.0281	−0.0010
19	−0.3698	1.2223	−1.5452	0.9006	−0.2577	0.0293	−0.0010
20	−0.7006	2.2013	−2.6177	1.4633	−0.3806	0.0441	−0.0010
21	−0.1180	0.3976	−0.5111	0.3106	−0.0944	0.0153	−0.0010
22	−0.4794	1.5831	−2.0142	1.2183	−0.3469	0.0391	−0.0010
23	−0.3859	1.2649	−1.5742	0.9162	−0.2497	0.0287	−0.0010
24	−0.5679	1.8266	−2.2197	1.2564	−0.3314	0.0360	−0.0010
25	−0.2271	0.7208	−0.8814	0.5260	−0.1633	0.0251	−0.0010
26	−0.2625	0.8845	−1.1517	0.7124	−0.2082	0.0254	−0.0010
27	−0.2724	0.9295	−1.2162	0.7457	−0.2109	0.0242	−0.0010
28	−0.5000	1.6587	−2.0858	1.2193	−0.3265	0.0344	−0.0010
29	−0.2628	0.8659	−1.0948	0.6587	−0.1922	0.0252	−0.0010
30	−0.1761	0.6540	−0.9515	0.6544	−0.2057	0.0248	−0.0010
31	−0.8975	2.7919	−3.5000	1.8469	−0.4870	0.0557	−0.0010
32	−0.2613	0.9013	−1.2143	0.7819	−0.2359	0.0282	−0.0010

Note: C_i denotes the ith coefficient of the polynomial

appropriate criterion for characterizing performance, we empirically code the residue for a 5 msec duration (50 samples at 10 kHz) by the use of 256 codewords. The type of codebook population is not a crucial factor from a perceptual point of view, experiments with Gaussian, sparse and ternary-value (-1,0,+1) codebooks have been reported to produce similar synthetic quality (Trancoso et al., 1990). However, since the probability density function of the unvoiced residue is nearly Gaussian, we still employ a Gaussian noise generator to establish the codewords. For each codeword, special formulations of its content are only for the purpose of reducing the computational effort, which is necessitated by the filtering process in codeword searching (Kleijn et al., 1990; Galand et al., 1992). This kind of computational burden can also be alleviated by other means such as the singular value decomposition (SVD), frequency domain and autocorrelation approaches (Trancoso and Atal, 1990). In our experiment, the autocorrelation approach is adopted to facilitate the computation. Some relevant details will be given in Chapter 4.

As mentioned in the previous section, samples for each codeword are drawn from a Gaussian noise generator, but we employ three schemes to established the codebook.

Scheme 1 (64 entries) — Each codeword contains 16 non-zero samples. The positions of non-zeros samples exhibit a uniform distribution from 1 to 50.

Scheme 2 (64 entries) — The conditions are the same in Group 1 except that 32 out of 50 samples are non-zero.

Scheme 3 (128 entries) — Every sample is taken from a Gaussian noise generator.

The sparse codewords in Schemes 1 and 2 are used to enhance the spiky nature of the residue so that the stochastic codebook can also be applied to synthesize the mixed sounds as well as plosives. This concept is very similar to that proposed by Kang and Everett (1985), who introduce a few spaced spikes into the unvoiced excitation in order to obtain satisfactory plosive sounds.

CHAPTER 4
SPEECH ANALYSIS/SYNTHESIS/EVALUATION

In Chapter 2, we focused on how to interpret the acoustic features of speech signals within the linear source-filter theory. The power of LP techniques for performing the feature extraction suggests that a high-quality LP synthesizer could be achieved if these features were appropriately modeled and accurately estimated. Hence, in Chapter 3 we discussed source modeling. The residue, known as the ideal source excitation, was simulated either by glottal impulses for voiced speech or innovation sequences for unvoiced speech. Both types of excitations were further formulated into two specific codebooks. The reader can envisage Chapter 2 as an anatomical study of the speech signals and Chapter 3 as an examination of the glottal source. The information obtained from these chapters can now assist us in deriving a synthesis model capable of producing high-quality natural-sounding speech.

In this chapter we present a new model which includes many improved features such as the interpolation of LP coefficients, turbulent noise and source-tract interaction. The parameters in this model are obtained by the analysis by synthesis procedure, in which the analysis denotes the process of estimating the parameters that characterize the speech signal and the synthesis denotes the process of replicating the speech signal by controlling and updating these parameters under the supervision of the speech production model. We will describe our methods and strategies in dealing with these issues. While the performance of this model is evaluated by judging its ability to produce natural speech, we also discuss the results of informal listening tests.

4.1 Analysis Scheme

The speech production model employed in this study is depicted in Figure 4–1. Except for the excitation source, the model retains the basic structure of the pitch-excited LP synthesizer. In addition to an all-pole filter, other parameters required by this model comprise a voicing decision, voiced/unvoiced gains, codeword indexes and Glottal Closure Instants (GCI's) for the voiced speech.

In general, a pitch synchronous approach is preferred for speech processing not only because it provides better formant trajectories (Krishnamurthy and Childers, 1986) but also because it facilitates the synthesis work. To implement such an approach, we need to locate every GCI accurate before computing the LP coefficients. The difficulties of identifying GCI's complicate the feasibility and reliability of the implementation, making the pitch synchronous analysis practically unattractive. Thus, we decide to use a frame-based method to compute the LP coefficients, but carry out the speech synthesis pitch synchronously after determining the pitch period.

Since the speech signal is sampled at 10 kHz, a linear predictor of 13th order is chosen to account for the spectral characteristics of the glottal source (3 poles) and vocal tract (10 poles). The filter coefficients along with the residue are derived concurrently using an orthogonal covariance method (Ning and Whiting, 1990), performed once per frame sequentially throughout the input speech. The frame size is 25 ms with an overlap of 5 ms between any two consecutive frames. For each frame, the LP gain is normalized by adjusting the power of the residue to that of the speech signal. The residue in the overlapped area is obtained by weighting the forward and backward overlapping sequences with decreasing and increasing trapezoidal windows respectively and adding them together.

$$e(i) = \frac{N+1-i}{N+1} e_f(i) + \frac{i}{N+1} e_b(i), \qquad i = 1, 2, 3, \ldots, N \qquad (4\text{–}1)$$

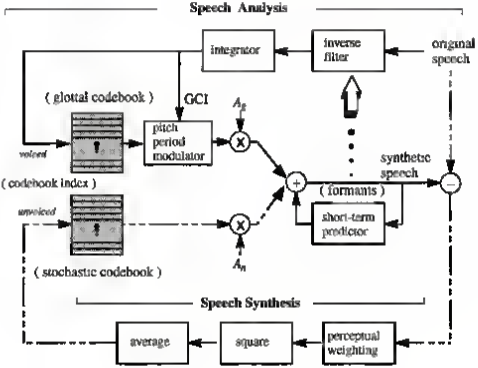


Figure 4–1 Proposed speech production model.

where $e_f(i)$, $e_b(i)$ denotes the forward and backward residue signals respectively. $e(n)$ is the resulting residue signal for the overlapped area of length N.

4.1.1 Orthogonal Covariance Method

Consider a digital signal with the following sequence, $\{s_1, s_2, \cdots s_p \cdots, s_{m+n}\}$. The linear prediction of the current sample is described as a linearly weighted summation of past samples, i.e.,

$$s_n = \sum_{k=1}^{m} a_k s_{n-k} + e_n \qquad (4\text{–}2)$$

where the a's are the coefficients of the LP predictor with order m, and the e's are the prediction errors. Expressing the equations above in a matrix form, we have

$$\begin{bmatrix} s_1 & s_2 & \cdots & s_m \\ s_2 & s_3 & \cdots & s_{1+m} \\ s_3 & s_4 & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ s_n & s_{n+1} & \cdots & s_{n+m} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} s_{m+1} \\ s_{m+2} \\ \vdots \\ s_{m+n+1} \end{bmatrix} - \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \qquad (4\text{–}3)$$

For the convenience of illustration, vector notation is employed in the following derivations. We define the S_k as the kth column vector of the matrix S, A as the vector of the LP coefficients, and E as the vector of prediction error. Thus, Eq. (4–3) becomes

$$[S_1\ S_2\ S_3\ \cdots\ S_m]A = S_{m+1} - E \qquad (4\text{–}4)$$

By assuming that the prediction error is negligible we may eliminate E and determine A by multiplying the pseudo inverse of S on both sides of Eq. (4–3). It may be shown that the obtained result is the same as that derived by a covariance method, because the error is minimized over a specified interval.

It can also be shown that a certain degree of efficiency could be gained by reformulating the foregoing computation as follows. Suppose we now decompose the vector S_{k+1} into k orthogonal vectors V_i's using the Gram Schmidt method. The set of the orthogonal vectors is

$$V_{k+1} = S_{k+1} - \sum_{i=1}^{k} c_k^i V_i \tag{4–5}$$

where $$c_k^i = \frac{S_i^t V_i}{V_i^t V_i}, \tag{4–6}$$

and the superscript t denotes the transpose operator. Arranging the orthogonal expansion in a matrix form, we obtain

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ c_2^1 & 1 & 0 & 0 & 0 & 0 \\ c_3^1 & c_3^2 & 1 & 0 & 0 & 0 \\ & & & & & \\ c_p^1 & c_p^2 & 0 & 1 & 0 & 0 \\ & & & & & \\ c_m^1 & c_m^2 & & & c_m^{m-1} & 1 \end{bmatrix} \begin{bmatrix} V_1^t \\ V_2^t \\ V_3^t \\ V_4^t \\ \\ V_p^t \\ \\ V_m^t \end{bmatrix} = \begin{bmatrix} S_1^t \\ S_2^t \\ S_3^t \\ S_4^t \\ \\ S_p^t \\ \\ S_m^t \end{bmatrix} \tag{4–7}$$

Through several algebraic manipulations, the row vector S_i^t on the right hand side of Eq (4–7) can be shown as

$$S_{p+1}^t = \sum_{k=1}^{p} c_{p+1}^k V_k^t + V_{p+1}^t = C_{p+1}^t [C]_{pxp}^{-1} \begin{bmatrix} S_1^t \\ S_2^t \\ S_3^t \\ \\ S_p^t \end{bmatrix} + V_{p+1}^t \tag{4–8}$$

where $\boldsymbol{C}$ is the matrix containing the c_k^i's on the right hand side of Eq (4–7) $[\boldsymbol{C}]_{pxp}$ is a upper left submatrix of the matrix $\boldsymbol{C}$ with a rank p, and C_{p+1}^t is the $(p+1)$th row vector of the matrix $\boldsymbol{C}$ with only the first p elements included. Compared to Eq (4–3), it is found

immediately that the coefficient vector A is equivalent to the term

$$A^t = C^t_{p+1}[C]^{-1}_{p\times p}, \tag{4-9}$$

and the vector V_{p+1} is just the estimated error (or the residue) of the pth order LP filter

A major advantage of using the orthogonal expansion is that the matrix inverse of C can be achieved by a back-substitution procedure (Ning and Whiting, 1990). Another important advantage emerges from the error vector V_{p+1}, which is known as a principal argument to determine the order in many methods. In speech processing, the importance of selecting a correct order can be explained in terms of formant characteristics. A filter with a lower order tends to disregard insignificant formants or to merge two adjacent ones, whereas a higher order filter raises the possibility of producing spurious formants. The resulting incorrect formants may lead to perceivable errors in both cases. Thus a variable-order predictor is always preferred, for it adapts the spectral variation of running speech. Apart from this reason, such a model can also reduce the transmission bandwidth if lower orders are frequently chosen.

There are two widely accepted order estimators, namely, the Akaike information criterion (AIC) (Akaike, 1974) and minimum description length (MDL) criterion (Schwarz, 1978) which can be obtained prior to estimating the LP coefficients. Two other methods proposed in the recent literature are the Predictive Least Squares (PLS) (Wax, 1988), and the Iterative Algorithm of Singular Value Decomposition (IASVD) (Konstantinides and Yao, 1988). A comparison of the performance of the four methods indicated that the IASVD had the highest success rate in order selection followed by MDL, AIC and PLS (Konstantinides, 1991). If we take into account computational efficiency, the MDL turns out to be a proper choice to work with the orthogonal covariance method. Eventually the selected order is the one that minimizes the MDL function given by

$$MDL(i) = N \times \ln(\ V_i^t V_i\ ) + i \times \ln(N) \qquad (4\text{–}10)$$

After we determine the optimal order p, the LP coefficients is then derived from Eq. (4–9).

4.1.2 V/U/S Classification

Because there are two types of excitation functions in the proposed model, the first step toward speech analysis is a voicing decision. The basic principles of our method are rather simple. If the energy of the underlying signal is below a specified value, the signal is classified as silence (Campbell and Thomas, 1966; Childers et al., 1989a). Otherwise, we examine its spectral tilt by calculating the first reflection coefficient. The signal is attributed to voiced speech if the first reflection coefficient is larger than 0.3. Unvoiced speech is the result when the previous two tests have failed.

Unlike other algorithms, the correct rate of classification is not strictly required because the incorrect decision does not lead to serious perceptual errors. For example, the misclassification between unvoiced and silence is not critical since both share the stochastic codebook and the quantization error in the silence can always be ignored. Also, the speech signal with a medium spectral tilt (e.g., the first reflection coefficient is around 0.3) often exhibits mixed characteristics of both types of excitations. Therefore, either voiced or unvoiced excitation is acceptable for synthesizing such a speech signal.

4.1.3 Identification of Glottal Closure Instant (GCI)

A reliable identification of the GCI is essential for codeword searching and speech synthesis since both are performed on a pitch period-by-pitch period basis. Procedures of the GCI identification algorithm can be summarized in two steps: (1) pitch estimation, and (2) peak picking. That is, we determine the location of glottal closure after estimating the pitch period.

It has long been noted that the sharp peaks in the residue signal generally coincide with the GCI for a wide variety of voiced sounds. Choosing the largest peak of the residue

signal for many voiced sounds is a trivial method for determining the GCI (Atal and Hanauer, 1971, Ananthapadmanabha and Yegnanarayana, 1979) For voices that are not rich in harmonic structure or that lack distinctive glottal closure may fail to have large peaks in the residue Furthermore the true peaks may be obscured by other spurious peaks due to background noise and modelling errors. Perhaps the easiest way to circumvent this drawback is to apply a lowpass filter to reduce the influences of the spurious peaks. However too much smoothing definitely decreases the sharpness of the real peak, so we can only eliminate the influence of noisy components to such an extent that the true peaks are not smeared out. To avoid the phase shift of the peaks, we perform the lowpass process by a zero-phase filter, i.e., by first passing the residue signal forward then running it back through the same filter (Oppenheim and Willsky, 1983) The Z-domain representation of the employed filter in our experiment is chosen to be

$$H(z) = \frac{1}{(1 - .9z^{-1})(1 - .7z^{-1})} \qquad (4\text{–}1)$$

Once the residue signal is lowpass filtered, a segment of 512 samples centered at the current frame is extracted using a hanning window. This windowed segment $s(n)$ is then transformed to a sequence $P_s(n)$ similar to the ceptrum by

$$P_s(n) = IFFT(|FFT[s(n)]|) \qquad (4\text{–}2)$$

where FFT and IFFT stand for the fast Fourier transformation and its inverse operation, respectively, and $|\bullet|$ denotes the magnitude.

Like the pitch estimation procedure outlined in the ceptrum method, we choose m as the pitch period if

$$P_s(m) > P_s(n) \qquad n = 25, 26, \ldots, 256. \qquad (4\text{–}3)$$

The value, m,could be a multiple of the real pitch period, however In our program, a simple

check is given as follows. We first look for the position l of the largest value within the range [25,m–25], i.e.,

$$P_s(l) > P_s(n) \qquad \text{for} \quad l \neq n;\ l,n = 25,\ldots,m-25. \tag{4–4}$$

If the following condition exists

$$P_s(l) > 7P_s(m), \tag{4–5}$$

then l is adopted as the new pitch period. Otherwise, the pitch period is remained as m.

After finding the pitch period, we begin with the search for the largest negative peak in the smoothed residue. Due to that the peak has been smeared by the zero-phase lowpass filter, we enhance the accuracy of peak picking by approximating the curve on both sides of the negative peak by two straight lines ranging from the peak value to one-third of this value (highlighted by the circled area in Figure 4–2(a)). The intersection of the two lines is chosen to be the first GCI. A small interval of samples (~4 5ms) around the first GCI is used as a template (as shown in a dashed box) to discriminate other peaks within the same frame. Peaks located before or after this GCI with approximately one pitch period are examined by computing the correlation between the template and the waveforms around the peaks. Positions that lead to largest correlations are then selected as other GCI's. This procedure continues until the searching range is out of the current frame by 50 samples

The overall computation above costs 2 FFT's and several comparisons. An economical approach for performing the whole process is to decimate the signal $s(n)$ by a factor of 2 and then to perform an interpolation on $P_s(n)$ to counteract such a decimation. Because of the lowpass filtering, the foregoing decimation can be carried out by choosing every other sample of $P_s(n)$ without causing serious aliasing.

A complete example of the GCI identification is illustrated in Figure 4–2(b). In this example, it appears that the GCI can be directly identified by picking the negative peaks of the lowpass filtered residue. But, as we mentioned earlier, the peaks in the residue signal

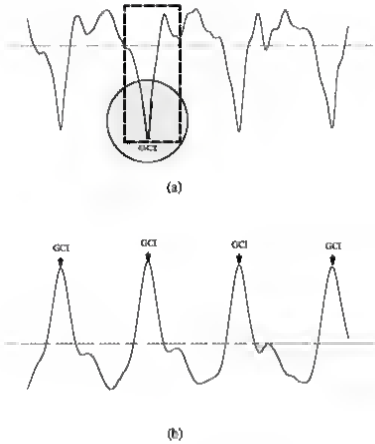


Figure 4-2 Illustration of GCI identification (a) lowpass filtered residue signal, (b) cross correlation for this residue signal with a template shown in the dashed box in (a)

are not always sharp and distinctive. The tedious work with regard to the smeared residue plays a role in helping to reduce potential errors. Similarly, the acquisition of GCI outside the current frame is redundant but necessary, because the extra information can be used to prevent erroneous GCI's in frame boundaries.

4.1.4 Codeword Searching

Depending on the voicing conditions, there are two different codebooks prepared to reconstruct the synthetic excitation. Although the basic idea of codeword searching for these two codebooks is the same, i.e., selecting a optimum codeword that achieves a minimum error subject to a distance metric, the individual implementations are somewhat different due to their intrinsic characteristics.

4.1.4.1 Voiced excitation, glottal codebook

The searching process for the optimal glottal codeword requires that the integrated residue and the polynomial waveform are of the same length. We assume the maximum allowable length for one pitch period to be 25.6 ms. Thus, if we encode every polynomial waveform with such a maximum length, then the integrated residue of one pitch period can always be interpolated to the maximum length using the FFT method. Taking advantage of the symmetry of the Fourier transformation, we compute the correlation coefficient, $\eta_m(i)$, between the ith polynomial waveform, $g_i(n)$, and the integrated residue of the mth period, $d_m(n)$, by

$$\eta_m(i) = \frac{real\left[\sum_{k=2}^{129} D_m(k)G_i'(k)\right]}{\left[\sum_{k=2}^{129} D_m(k)D_m'(k)\right]^{1/2}\left[\sum_{k=2}^{129} G_i(k)G_i'(k)\right]^{1/2}} \qquad (4\text{-}6)$$

$$= \frac{real\left[\sum_{k=2}^{\lceil l/2 \rceil} D_m(k) G_i'(k)\right]}{\left[\sum_{k=2}^{\lceil l/2 \rceil} D_m(k) D_m'(k)\right]^{1/2} \left[\sum_{k=2}^{\lceil l/2 \rceil} G_i(k) G_i'(k)\right]^{1/2}} \qquad (4\text{–}6)$$

where $D_m(k)$ is the FFT sequence of the interpolated $d_m(n)$, $G_i(k)$ is the FFT sequences of $g_i(n)$, $\lceil * \rceil$ denotes the ceiling function and $\lceil ' \rceil$ denotes the complex conjugate. It is noted that the mean values of $g_i(n)$ and $d_m(n)$ are zero and, therefore, play no role in computing the correlation coefficient. We reflect this consequence by skipping the d.c term during the multiplication of two FFT sequences. The second equality in Eq. (4–6) is due to the fact that the interpolated FFT sequence of $D_m(k)$ is zero when $k \geq \lceil l/2 \rceil$. The spectrally weighting filter, which is commonly used in the CELP coder, does not participate in the equation above. This is because our distance measure is applied to the integrated residue which emphasizes only on the glottal phase characteristics at the low frequency region.

Since the glottal waveform varies relatively slowly compared to the changes of the vocal tract transfer function, one codeword index is found to be enough to describe the glottal excitation for each voiced frame. We further define the cumulative similarity function, $H(i)$, as the sum of $\eta_m(i)$ along one frame.

$$H(i) = \sum_{m=1}^{N} \eta_m(i) \qquad (4\text{–}7)$$

where N is the total number of the pitch periods in this frame. The codeword that leads to the maximum cumulative similarity is chosen as the representative for the entire frame.

4.1.4.2 Unvoiced excitation: stochastic codebook

In this category, the codeword searching method is that commonly used in CELP coders. The remainder of this section provides a brief discussion of the CELP algorithm and the autocorrelation method that achieves a fast codeword searching.

4.1.4.2.1 CELP algorithm

The CELP algorithm was first proposed by Schroeder and Atal in 1984. A rapid development standardized this algorithm in the late 1980's. The CELP coder represents a breakthrough in speech coding for it encodes speech signals at a rate as low as 4.8 Kb/s but still produces a satisfactory quality. The basic concept for the class of CELP coders can be viewed as a vector quantization technique, which passes a finite set of candidate vectors through an all-pole predictor filter and then selects the one giving a best match subject to a specific error criterion. However, the research that led to the development of CELP coders seemed to follow another path emerging from the MPLP coder, in which the excitation consists of a few pulses per frame regardless of whether the speech is voiced or unvoiced. The locations and amplitudes of these pulses are determined by minimizing a subjective error between the original and synthetic speech signals. The relationship between the CELP and MPLP coders can be understood by considering the multipulse excitation as a deterministic codebook consisting of innovation sequences (or codewords), each consisting of an single impulse with a different delay. Hence, searching for an optimum pulse location across the analysis frame is equivalent to searching through a set of ensembles.

In the primitive CELP coder (Figure 4–3), the speech signal, $s(n)$, is analyzed in blocks of N samples. For each block, the synthetic speech signal is derived by filtering every innovation sequence stored in a codebook into two recursive filters (long term and short term) with a proper scaling factor. An error signal is then formed by comparing the synthetic speech to the original one. Through an exhaustive search over the entire codebook,

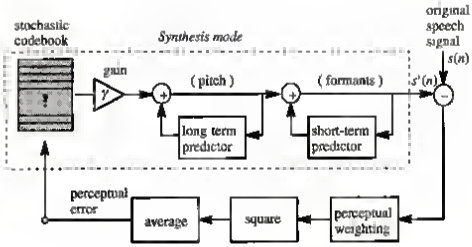


Figure 4–3. Block diagram of the CELP vocoder

the innovation sequence (along with an appropriate scaling factor) that produces the minimum mean-squared subjective error is selected to reconstruct the synthetic excitation.

The short-term predictor in the CELP coder is the well-known LP filter. The long-term predictor is an extra stage used to enhance the periodicity of the synthetic speech by exploiting the similarity across consecutive pitch periods, and has been applied in open-loop and closed-loop form. In the former case the long-term predictor is directly derived from the residue obtained by inverse filtering the original speech, while in the latter case the optimal long-term predictor is computed based on an analysis-by-synthesis procedure. Although the analysis-by-synthesis procedure does not provide much improvement of speech quality over the open-loop procedure, it spawns the concept of the "adaptive codebook" or "self-excite" model, in which the codebook entry is defined as the application of a moving window to the recent past excitation. More precisely, each codeword is a shifted version of the previous one with one new sample changed at the end. The conceptual structure of the adaptive codebook is illustrated in Figure 4–4. As seen in this figure, the function of the pitch predictor is replaced by the adaptive codebook. Owing to the dependency of the neighboring codewords, together with a relaxed error criterion that provides an even weighting to the codewords, fast algorithms have been exploited to reduce the inherently high computational complexity of closed loop procedure.

Following the formalism given by Trancoso and Atal (1990), we now use the matrix notation as well as vector notation to illustrate the analysis-by-synthesis procedure for codeword searching. Given a codebook of L sequences $C_n^{(k)}$ (k=1,2, . . . , L) each of length N, the filtering operation for an innovation sequence by the long- and short term filters can be carried out by convolving the innovation sequence with the combined impulse response of these two filters. Written in matrix form, the filter output for the kth codeword can be represented by

$$y^{(k)} = x^{(k)}Hc^{(k)} \qquad (4\text{–}8)$$

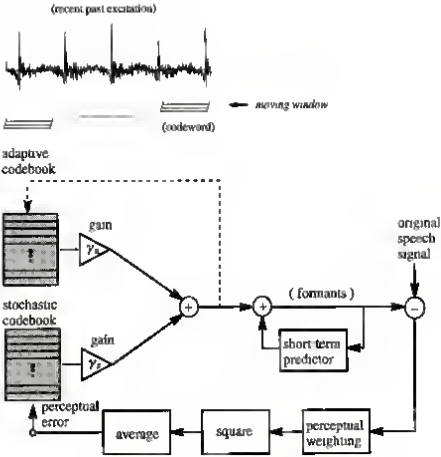


Figure 4–4. Illustration of the adaptive codebook.

where $\gamma^{(k)}$ is the scaling factor for the kth codeword, H is an $N \times N$ matrix with the element in the mth row and the nth column given by the $(m-n)$th sample of the unit impulse response of the filter, and $c^{(k)}$ is a N dimensional vector with its nth component given by $c_n{}^{(k)}$. Since $h_n = 0$ for $n<0$, the matrix H can be shown as

$$H = \begin{bmatrix} h_0 & 0 & & & 0 \\ h_1 & h_0 & & & \\ & h_1 & & & \\ & & & & \\ & & & h_0 & 0 \\ h_{N-1} & h_{N-2} & & h_1 & h_0 \end{bmatrix} \tag{4–9}$$

Let us define x to be the desired signal with its nth component given by x_n, of which the memory contribution carried over previous frames has been removed since the filter memory plays no role in the search procedure. The total squared error $E^{(k)}$, representing the difference between the desired vector x and the vector $\gamma^{(k)}$, is defined as

$$E^{(k)} = \| x - \gamma^{(k)} H c^{(k)} \|^2, \tag{4–10}$$

where $\| \|^2$ indicates the squared norm of the underlying vector. The optimum scale factor $\gamma^{(k)}$ that minimizes $E^{(k)}$ is determined by setting $\partial E^{(k)} / \partial \gamma^{(k)} = 0$, yielding

$$\gamma^{(k)} = \frac{x^t H c^{(k)}}{\| H c^{(k)} \|^2}, \tag{4–11}$$

and the error becomes

$$E^{(k)} = x^t x - \frac{[x^t H c^{(k)}]^2}{c^{(k)t} H^t H c^{(k)}} = \| x \|^2 - \frac{[x^t H c^{(k)}]^2}{\| H c^{(k)} \|^2} \tag{4–12}$$

The best codeword is obtained by selecting the index k in a exhaustive search for which the error $E^{(k)}$ is minimum or, equivalently, the second term on the right hand side of Eq. (4–12) is maximum.

In principle, the error derived above spans over the entire spectrum of the synthetic speech. Due to auditory masking, the error in the high energy regions is masked by the speech signal, suggesting that the error should be concentrated in the formant regions to reduce perceptual distortion. This idea can be easily accomplished by the use of a weighting filter $W(z)$ that attenuates the frequencies where the error is perceptually less important and amplifies those frequencies where the error is perceptually more important.

$$W(z) = \frac{1 - \sum_{k=1}^{p} a_k \zeta_1^k z^{-k}}{1 - \sum_{k=1}^{p} a_k \zeta_2^k z^{-k}} \qquad (4\text{–}13)$$

where $0 \le \zeta_2 < \zeta_1 \le 1$, and a_k is the LP coefficients. If ζ_1 is set to be unity, ζ_2 in the range $0.6 \le \zeta_2 \le 0.9$ gives similar subjective results in informal listening tests (Rose and Barnwell, 1990).

Referring to Eq (4–12), the computation required in the codeword searching contains only two terms, namely, a cross correlation term between vectors $x^t H$ and $c^{(k)}$, and an energy term corresponding to the filtered output $Hc^{(k)}$ of each codeword. The energy term is computationally complicated if the matrix multiplication $Hc^{(k)}$ is directly performed. Fortunately, many methods have been proposed for avoiding the time-consuming matrix multiplication. We will now discuss the autocorrelation method, which is very efficient for fully populated codebooks and, therefore, an excellent choice for our stochastic codebook.

4.1.4.2.2 Autocorrelation method

Let us first consider the energy term in the second part on the right side of Eq (4–12). Recall that we already dropped the long term predictor in our model. The h_m only represents the sequence of the impulse response of the short-term predictor filter. We rewrite the energy term in scalar notation,

$$\| Hc^{(k)} \|^2 = \sum_{n=1}^{N} \left[\sum_{i=1}^{N} h_{n-i} c^{(k)} \right]^2 \tag{4–14}$$

Making use of the fact that the sum of the squares of the convolution of two sequences equals the cross correlation of the autocorrelations of these two sequences, Eq. (4–14) can be simplified as

$$\begin{aligned} \| Hc^{(k)} \|^2 &= \sum_{i=1}^{N} \sum_{j=1}^{N} c_i^{(k)} c_j^{(k)} \sum_{n=1}^{N} h_{n-i} h_{n-j} \\ &= R_h(0) R_c^{(k)}(0) + 2 \sum_{i=1}^{N-1} R_h(i) R_c^{(k)}(i) \end{aligned} \tag{4–15}$$

where

$$R_h(i) = \sum_{n=0}^{N-1-i} h_n h_{n+1} \tag{4–16}$$

and

$$R_c^{(k)}(i) = \sum_{n=0}^{N-1-i} c_n^{(k)} c_{n+i}^{(k)} \tag{4–17}$$

For Eq. (4–15) to be held, the convolution cannot be truncated, requiring that the impulse response of the synthesis filter is effectively zero beyond the N sample. In most circumstances this requirement will be satisfied after imposing the spectral weighting filter. If we further define the cross-correlation between h_n and x_n by

$$R_x(i) = \sum_{n=0}^{N-1-i} x_n h_{n-i} \tag{4–18}$$

Eqs. (4–11) and (4–12) are transformed to

$$\gamma^{(k)} = \frac{\left[\sum_{i=0}^{N-1} R_x(i)c_i^{(k)}\right]}{R_h(0)R_c^{(k)}(0) + 2\sum_{i=1}^{N-1} R_c(i)R_h^{(k)}(i)} \tag{4-19}$$

and

$$E^{(k)} = \| x \|^2 - \frac{\left[\sum_{i=0}^{N-1} R_x(i)c_i^{(k)}\right]^2}{R_h(0)R_c^{(k)}(0) + 2\sum_{i=1}^{N-1} R_c(i)R_h^{(k)}(i)}, \tag{4-20}$$

respectively

From above derivation, it is easily seen that this method contributes substantial savings in computation time. The energy term can now be computed with just N multiplications per codeword However, the price we have to pay is the storage of an additional codebook with the autocorrelation coefficients of the original codewords.

4.2 Synthesis Scheme

Speech synthesis is the procedure of reconstructing speech signals by controlling and updating the parameters of a speech production model estimated in speech analysis. The synthesis of unvoiced speech is straightforward and can be easily accomplished by exciting the time-varying all pole filter with the gain adjusted innovation sequence sequentially On the other hand, the synthesis of voiced speech is rather complicated because we have to construct the synthetic excitation from the gross features of glottal phases Therefore, most of this section is focused on the synthesis schemes for voiced speech.

Despite many control parameters for voiced speech were estimated on a frame-by-frame basis, the corresponding synthesis can still be carried out pitch synchronously provided that the control parameters are properly interpolated for each pitch

period. In this section, we start from the discussion of the interpolation with respect to the glottal phase and LP coefficients. Then, we present a method for eliminating the spectral tilt of the glottal pulse. Effects of vocal noise and source-tract interaction are discussed subsequently. Finally, a complete procedure for generating a glottal impulse is given.

4.2.1 Interpolation of Glottal Phase

As mentioned earlier, only one codeword is employed to indicate the glottal phase characteristics for each frame. Although large discrepancies may occur between any two adjacent pitch periods in one frame, a progressive alteration of the glottal pulse still sounds reasonable because such discrepancies are already reflected in the codewords of different frames. This, however, results in the discontinuities of the glottal phase characteristics at the frame boundaries. Since the glottal phase is manifested as a sixth order polynomial, we therefore apply a lowpass filter to eliminate the rapid changes of the polynomial coefficients as follows.

(1) IIR filter:

$$\hat{P}^i(k) = (1-\alpha)\hat{P}^{i-1}(k) + \alpha P_m(k) \tag{4-21}$$

(2) FIR filter:

$$\begin{cases} \hat{P}^i(k) = \alpha P_{m-1}(k) + \beta P_m(k) + \gamma P_{m+1}(k) \\ \alpha + \beta + \gamma = 1 \end{cases} \tag{4-22}$$

where $\hat{P}^i(k)$ is the polynomial for the ith pitch period, and $P_{m-1}(k)$, $P_m(k)$ and $P_{m+1}(k)$ are the polynomials for the previous, current and next frames, respectively. In our program, an IIR filter with the value of $\alpha = 0.5$ is used since it works well in our experiments. We recall that the resulting polynomial has to satisfy the three constraints specified in Section 3.2.1. Therefore, the condition where the sums of the coefficients on the right hand sides of Eqs. (4–21) and (4–22) are unities is to comply with the first constraint. However, no extra consideration is necessary for the other two constraints.

4.2.2 Interpolation of LP Coefficients

As in the case of the glottal phase characteristics, the LP coefficient extracted from the frame-based method may exhibits unreasonable discontinuities at frame boundaries. A simple solution will be to linearly interpolate the LP coefficients. However, synthetic speech produced by this method may sound too smooth for speech segments with a rapid spectral transition. The plosives are typical examples that suffer the drawback of the linear interpolation. This suggests that the interpolation of the LP coefficients should be "piece-wise continuous." Thus, we adopt a quadratic weighting function w_i to interpolate the LP coefficients:

$$w_i = \frac{\dfrac{1}{\left(|n_b - l_f/2 - i\, l_f| + |n_e - l_f/2 - i\, l_f|\right)^2}}{\displaystyle\sum_{j=-1}^{1} \frac{1}{\left(|n_b - l_f/2 - j\, l_f| + |n_e - l_f/2 - j\, l_f|\right)^2}}, \qquad i = -1, 0, 1 \quad (4\text{–}23)$$

where $|\cdot|$ denotes the absolute value, n_b and n_e denotes the positions of the beginning and ending points of the current pitch period, and l_f is the number of samples in each frame. The vector of the interpolated LP coefficients, A_{new}, is obtained by

$$A_{new} = A_{-1} w_{-1} + A_0 w_0 + A_1 w_1 \quad (4\text{–}24)$$

where A_{-1}, A_0, A_1 are the LP coefficient vectors of the pervious, current and next frames, respectively. Figure 4–5 illustrates the linear and quadratic interpolations for an arbitrary coefficient. What we mean by "step-wise linear" is clearly delineated by the quadratically interpolated curve.

One of the disadvantages of such an LP interpolation is that it occasionally moves poles outside the unit circle, implying that we have to reflect these outside poles into the unit circle in order to stabilize the synthesis filter. However, we do not consider this to be a serious problem since the interpolation can also be done using reflection coefficients, autocorrelation functions, cross-sectional areas, for all of which the stability criterion is

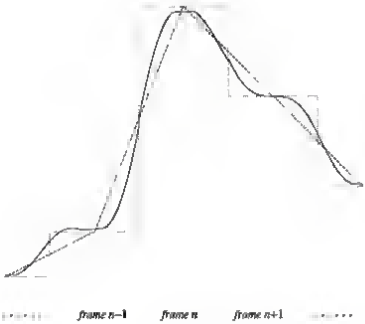


Figure 4–5. Interpolation with respect to one of the LP coefficients along several frames. (dotted line: without interpolation; solid line: with linear interpolation, dashed line: with proposed quadratic interpolation)

satisfied. Moreover, the proposed LP interpolation conflicts with the use of a variable-order filter. Apparently, the deadlock has to be resolved by inventing a transformation that is capable of performing interpolation with a different dimension. Unfortunately, we do not have an appropriate method for solving this problem. For this reason, a fixed-order filter still serves in this study.

4.2.3 Spectral Flatness

In order to meet the spectral specification of the residue, any source model for the LP synthesizer should have a flat spectrum. Our method for achieve the spectral flatness of the glottal excitation is inspired by the appearance of the integrated residue, in which the pulse swings around the glottal closure contribute most of the high frequency energy. Our formulation is given as follows:

First, we modify the third sample, $g(3)$, of the modeled polynomial waveform of the integrated residue so as to ensure the existence of a sharp pulse.

$$g(3) = (1 + \max\{|g(n)| \; n = 1, 2, \ldots, \lfloor l/4 \rfloor\}) * .92 - 1 \qquad (4\text{–}25)$$

where $g(n)$ represents the modeled integrated residue with the pitch period of l samples. Next, the fourth and sixth samples are changed to be

$$g(4) = \frac{g(3) + g(4)}{2} \qquad (4\text{–}26)$$

and

$$g(6) = \frac{g(3) + g(6)}{2}, \qquad (4\text{–}27)$$

(using the new value of $g(3)$) so that energy at the middle frequencies is enhanced. The excitation pulse is obtained by taking the differentiation of $g(n)$. Finally, a first order inverse filter is applied to remove the spectral tilt of the excitation pulse. The resultant excitation is denoted as the glottal impulse, which will be frequently seen in the rest of this dissertation.

4.2.4 Effect of Vocal Noise

Vocal noise is important for synthesizing breathy and female voices (Klatt, 1987, Pinto et al., 1989, Klatt and Klatt, 1990; Childers and Lee, 1991) In Chapter 2, we have shown that the extracted noise exhibits the following two features. First, the noise, consisting of the turbulent noise and modeling errors, had a flat spectrum. Second, the magnitude of such noise near the glottal closure is higher than that at the other places.

As we also pointed out, part of the vocal noise possibly resulted from the phase misalignment. In order to verify this possibility, we decide to run a simulation using a noise source that has a larger amplitude around the glottal closure. The noise is produced by modulating uniformly distributed white noise with a Gaussian window given by

$$W_g(n) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(2n/l)^2}{2\sigma^2}\right) + B, \quad -\lfloor l/2 \rfloor \le n \le \lfloor l/2 \rfloor \quad (4\text{–}28)$$

where l is the pitch period, $\lfloor * \rfloor$ denotes the floor function. By referring to the measurements in Chapter 2, we choose σ as 0.25 and B as 0.5 to approximate the amplitude modulation of the vocal noise for normal male subjects. While adding this noise to the synthetic excitation, the amplitude of the vocal noise is adjusted to achieve a Signal-to-Noise Ratio (SNR) of 25dB.

For the sake of comparison, we also test another type of vocal noise with a constant modulation, of which the level is measured at the middle between two glottal closure instants. Furthermore, we adopt a 60% duty cycle starting at the maximum glottal closure since it was preferred in a listening evaluation (Childers and Lee, 1991) The SNR is modified as 28 dB to meet the measured level.

4.2.5 Source-tract Interaction

Source-tract interaction has been conjectured to be important for synthesizing high-quality, natural-sounding speech (Yea et al.,1983, Allen and Strong, 1985) In order

to develop a comprehensive source model for speech synthesis, this particular effect cannot be neglected. The interaction between the source and tract can be achieved either by using a vocal system to control the glottal impedance or by incorporating an equivalent effect into a source model. Our approach falls in the latter category.

Two major effects in the glottal flow, namely, skewness and formant ripples, result from the source-tract interaction (Rothenberg, 1981; Fant and Ananthapadmanabha, 1982). The skewness, in general, varies at a relatively slow rate. Our glottal impulse model is expected to imitate the skewness with adequate precision. The formant ripples, on the contrary, are too subtle to be depicted by this model. Due to the lack of accurate estimation, we only present methods for imitating the formant ripples rather than direct modeling.

Since the ripple effect is associated with an increase in the formant bandwidth during the glottal open phase, similar results can be achieved by moving the poles of the LP filter inward or outward as we did for the spectral weighting filter in Section 4.1.4.2. The damping of an all-pole filter can be controlled by multiplying the corresponding coefficients, a_k's, by the powers of a factor α, i.e., $\hat{a}_k = a_k \alpha^k$ (Viswanathan and Makhoul, 1975; Tohkura et al., 1978). A value of α smaller than 1 will move the poles toward the origin and broaden the bandwidths of the poles. However, the opposite statement is not necessarily true when α is greater than 1. This is because the bandwidths are reduced only if the poles are moved closer to the unit circle.

One possible implementation for the ripple effect is to use two sets of LP coefficients to simulate the damping factors for two different glottal phases. We apply the normal LP coefficients during the glottal closed phase and switch to the LP coefficients with a larger damping (i.e., α<1) when the vocal folds are open. As illustrated in Figure 4-6(b), the damping of the formant energy in the synthesized vowel */i/* is quite evident.

In case the true bandwidths are narrower than those estimated by LP analysis (Woog, 1980), we instead decrease the damping of the close phase to compensate for the reduced bandwidths. Of course, this work can be accomplished by using a filter with $\alpha > 1$, but the

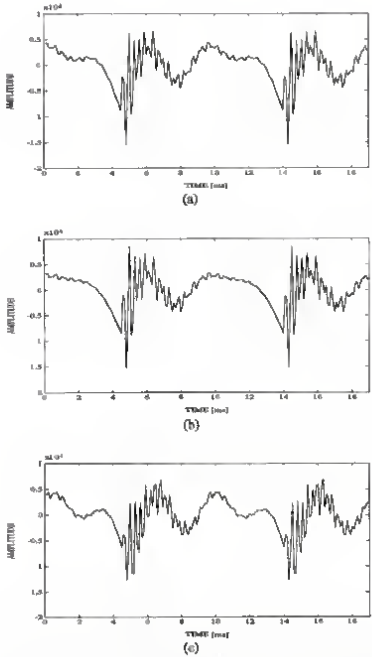


Figure 4–6 Synthetic vowel /i/ with. (a) original damping, (b) increased damping during the glottal open phase, (c) decreased damping during the glottal closed phase

implementation of such a filter requires a priori knowledge about the locations of the poles. This calls for a root-solving routine which is computationally expensive and very sensitive to the quantization errors. Thus, an alternative solution that we adopt in this research is to employ a filter, $W(z)$, to modify the excitation during the glottal closed phase so that the synthesized speech has a similar ripple effect. The filter is given by

$$W(z) = \frac{(1-\alpha z^{-1})(1-\sum_{k=1}^{p} a_k \beta^k z^{-k})}{(1-\beta z^{-1})(1-\sum_{k=1}^{p} a_k \alpha^k z^{-k})} \qquad (4\text{–}29)$$

where the values for α and β are 0.8 and 0.7, respectively

Although $W(z)$ takes effect only at the glottal closed phase, the filter memory carried over from the previous frame has to be taken into account. To reflect the contribution of filter memory, we impose a virtual constraint that the glottal impulses for any two consecutive periods are the same. Thus, filter memory can always be derived from the present glottal impulse instead of referring to the filtering results of the previous frame. By taking advantages that the duration of the glottal closed phase is usually less than one-half of the pitch period for modal voices and that the memory depth is limited by the filter order, we can perform the memory recovery process together with the ripple effect by filtering a circularly shifted glottal impulse with the glottal closure in the center (see Figure 4–7(5)). In this manner, the filter regains its memory during the glottal open phase and distributes the memory influence to the excitation approximately during the glottal closed phase.

To smooth the above process, we apply a hanning window to this circularly shifted glottal impulse. The windowed component is fed into the filter $W(z)$ to yield the intended damping. On the other hand, the remaining part (obtained by subtracting the windowed component from the excitation pulse) is kept unchanged throughout the course of filtering operation. Because $W(z)$ is an IIR filter, we append zeros to the windowed component to form a length of one and half pitch periods such that the filter is able to release its energy

sufficiently. The released energy is absorbed by adding the filtered component in the additional half period to that in the first half period. The glottal impulse with the source-tract interaction feature is formed by accumulating the filtered and remaining components together. Figure 4–6(c) shows the synthesized data produced by using such a glottal impulse. In this figure, the damping of the presented speech signal is obviously reduced in the glottal closed phase.

4.2.6 Generation of Glottal Impulse

To generate a glottal impulse, the recommended order for the implementation of the spectral flatness, vocal noise and source-tract interaction is given as follow:

1 generate a waveform $g(n)$ from the polynomial model.
2 modify $g(3)$, $g(4)$ and $g(6)$ by using Eqs. (4–25), (4–26) and (4–27), respectively
3 add white noise
4. differentiate the waveform to yield an excitation pulse.
5 simulate the source tract interaction.
6 remove the spectral tilt of the excitation pulse by inverse filtering.

Figure 4–7 illustrates the procedures above.

4.3 Gain Determination

In a speech production model the gain is a function to control the power transition along an utterance. Although the gain is an important factor affecting synthesis quality, unfortunately it has not received much attention from researchers. In this section we discuss methods for calculating the gain and their influences on synthetic speech.

4.3.1 Gain of Voiced Excitation, A_g

For a source-filter speech production model (Fant, 1960), the filter output can be decomposed into two components: one results from the excitation, $A_g u(n)$, and the other

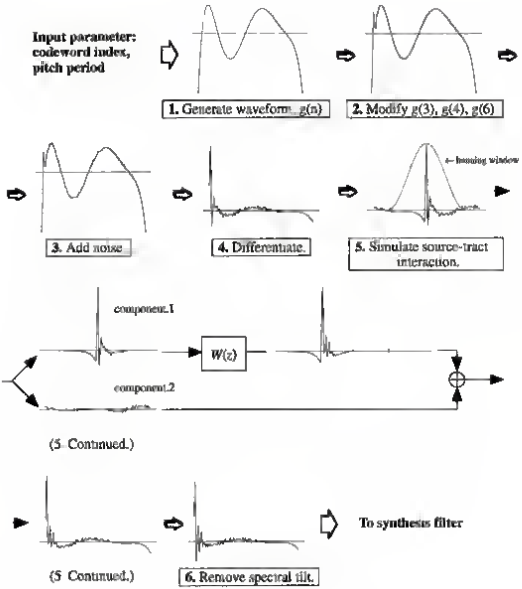


Figure 4–7. Procedures of generating a glottal impulse

results from the filter memory, $q(n)$. According to such a structure, a superposition method is often adopted for speech synthesis in to the avoid transient discontinuities at the boundaries of pitch periods (Verhelst and Nilens, 1986). That is, for each pitch period there are two synthesis filters employed: the one holding the previous LP coefficients is in charge of the memory contribution, and the other possessing the new LP coefficients is responsible for the current excitation.

Suppose we insist that the power of the filter output has to equal that of the original signal. Given a speech segment $s(n)$ of M samples with power P_r,

$$P_r = \frac{1}{M}\sum_{n=1}^{M} s^2(n) \tag{4–30}$$

Atal and Hanauer (1971) derived the gain A_g, by solving the following equation directly

$$P_r = \frac{1}{M}\sum_{n=1}^{M} (q(n) + A_g u(n))^2 \tag{4–31}$$

In case A_g was negative or complex, they set A_g as zero. The reason for causing such a zero setting is because the power contributed by the filter memory is too much. It appears that the zero setting is just a strategy to let the filter memory die out so that the gain can resume its function. Although this zero setting seems the only solution in order to make the synthesis implementable, it definitely destroys the pitch harmonics of the synthetic speech.

Tohkura et al. (1978) suggested that the memory contribution was negligible when the filter response was sufficiently damped. Thus, after increasing the damping factor of the synthesis filter, they computed the gain without considering the filter memory, i.e.,

$$P_r = \frac{1}{M}\sum_{n=1}^{M} \left(A_g u(n)\right)^2 \tag{4–32}$$

In consequence, the elimination of the zero setting is at the price of possible errors due to the ignorance of filter memory.

Makhoul (1975), on the other hand, computed the gain on the level of the driving function. By assuming that the excitation was either an impulse train for voiced speech or white noise for unvoiced speech, both of which have unity power, he computed the gain by estimating the power of the residue:

$$A_g = \left[R(0) - \sum_{k=1}^{p} a_k R(k) \right]^{1/2} \tag{4-33}$$

where $R(k)$ is the autocorrelation function of the analyzed speech signal and a_k's are the LP coefficients. Because the gain is a by-product of the LP analysis, this method seems very elegant and straightforward. However, it leads to the following problem. Unlike the impulse, the residue signal does not have a absolutely flat spectrum. Small mismatches between the impulse and the residue at low frequencies may be amplified after imposing the synthesis filter. Often times, the resultant errors are manifested as energy fluctuations in the synthetic speech, and a warble-like quality will be perceived.

In CELP and MPLP coders, the part attributed to the filter memory is first removed form the analyzed speech signal (Trancoso and Atal, 1990; Rose and Barnwell, 1990). The gain is determined by the cross correlation between the spectrally weighted speech signal and the weighted filter response of a given excitation function. This approach can automatically compensate for the resultant error when the synthesized speech signal does not match the original very well. Unfortunately, such an approach does not suit our model because the best fit of the glottal impulse may still result in a large discrepancy between the original and synthetic speech signals.

From the above discussions, we see that the gain is used to regulate the amplitude of the filter response of a given excitation. The segmental power P_r contains the information pertaining to the amplitude of the speech signal. To avoid the above-mention drawbacks, we demonstrate a method below to retrieve the gain from P_r.

It is noted that the speech waveforms in many adjacent pitch periods are very similar, suggesting that the initial and final filter memory are nearly equal in most cases. Thus, the filter memory can be approximated via the filtering operation with the use of the same excitation. If the number of pitch period, m, of the underlying excitation be large enough (say m=5) the filter memory contributed before the first period are negligible. Therefore, the zero-startup filter response in the last period can be regarded as a complete filter output. Two examples of such a filtering operation are illustrated in Figure 4–8 The gain A_g for the excitation pulse is then calculated by

$$A_g = \sqrt{\frac{P_r}{\frac{1}{l}\sum_{n=1}^{l} s_f^2(n)}} \tag{4–34}$$

where $s_f(k)$ is the resultant filter output within the mth pitch period, and l is the length of the pitch period.

The derivation described above needs a large amount of filtering operations. An algorithm presented in the following is provided to alleviate the computational burden. Notice that the filter memory from the past frame is always accessible during the speech synthesis We can simulate the foregoing filtering process by referring to the actual filter memory Suppose the filter is implemented using a direct-I form structure. The filter memory is, therefore, represented by the ending samples of the previous frame. As seen more obviously in Figure 4–8(b), it is the deviation of the filter memory that retains the similarity across all the periods. Based upon this observation, we separate the filter memory, $s(-k)$, into two parts the mean value, $s_m(-k)$, and the deviation, $s_d(-k)$:

$$s_m(-k) = \frac{1}{p}\sum_{k=1}^{p} s(-k) \qquad \text{for } k = 1, 2, \cdots, p, \tag{4–35}$$

$$s_d(-k) = s(-k) - s_m(-k) \quad \text{for } k = 1, 2, \cdots, p \tag{4–36}$$

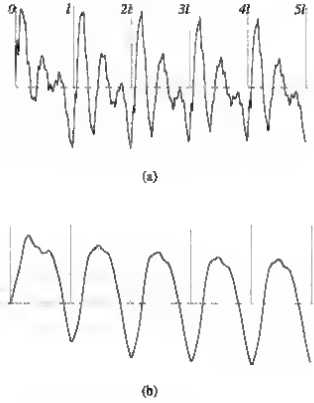


Figure 4–8 Zero-startup responses for two excitations of five consecutive cycles

We simulate the filtering process by means of an iterative procedure, which is governed by a constant variance of $S_d(-k)$. That is, in each iteration we calibrate the amplitude of the filter memory according to the mean value and deviation of the previous results by

$$s^{(i)}(n) = u(n) + \lambda\left[\sum_{k=1}^{p} a_k(U(n+p) - U(n))\right] + \mu\left[\sum_{k=1}^{p} a_k s_d^{(i-1)}(n-k)\right],$$
$$(n = 1, 2, ..., l) \tag{4-37}$$

where $u(n)$ is the filter response of the current excitation, $U(n)$ is the step function, and the superscript i denotes the iteration number. During each iteration, the scaling factors λ and μ are adjusted by

$$\lambda = \frac{1}{p}\sum_{k=1}^{p} s^{(i)}(-k), \tag{4-38}$$

$$\mu = \left[\frac{\sum_{k=1}^{p}(s_d^{(i)}(-k))^2}{\sum_{k=1}^{p}(s_d(-k))^2}\right]^{1/2} \tag{4-39}$$

We start with the iteration from the zero-startup response using the selected glottal impulse. After proceeding this iterative procedure several times, the resultant signal will approach to the one derived by filtering consecutive glottal impulses. The gain A_g is calculated as

$$A_g = \sqrt{\frac{P_r}{\frac{1}{l}\sum_{n=1}^{l}(s^{(i)}(n))^2}} \tag{4-40}$$

The obtained A_g's sometime exhibit large variations among adjacent pitch periods. This may cause perceivable energy fluctuation for synthetic speech. Unfortunately, a smoothing procedure based on A_g's cannot remedy such a defect because it does not take the filter gain into account. We solve the problem by introducing a new variable, ρ, which is

defined as the square root of the proportion of the power emerging from the current excitation

$$\varrho = \sqrt{\sum_{n=1}^{l} (u(n))^2 / \sum_{n=1}^{l} (s^{(0)}(n))^2} \qquad (4\text{–}41)$$

Then it is reasonable for us to argue that the obtained ϱ's vary slowly during voiced speech. We therefore apply a first order IIR lowpass filter, $0.3/(1 - 0.7z^{-1})$, to ϱ's to prevent occasional large power excursions across pitch periods. The final result of the gain A_g is determined by

$$A_g = \sqrt{\frac{P_r}{\frac{1}{l}\sum_{n=1}^{l} (u(n))^2}}\varrho \qquad (4\text{–}42)$$

Eventually, with the use of the proposed algorithm, the time-consuming convolution necessitated by the filtering operation is replaced by multiplications and additions. More important, this algorithm prevents the drawbacks occurred in other methods.

4.3.2 Gain of Unvoiced Excitation A_n

As mentioned in Section 4.1.4.2, we have adopted the CELP algorithm to reconstruct the unvoiced excitation. The gain A_n is the scale factor γ corresponding to the optimum codeword. While the optimal γ provides the minimum error, it often lowers the power intensity of synthetic speech. Hence, as recommended by Zinser and Koch (1989), the gain A_n is better replaced by a power match between the input signal $x(n)$ (signal with the memory contribution removed) and the filtered response of the synthetic excitation $y(n)$, i.e.,

$$A_n = \left[\frac{\sum_{j=1}^{m} x^2(j)}{\sum_{j=1}^{m} y^2(j)} \right]^{1/2} \qquad (4\text{–}43)$$

where m is the length of the subframe. In our case, $m = 50$

4.3.3 Voicing Transition

Similar to many LP synthesizers that use a multi-mode excitation source, an inappropriate voicing decision may lead to the deterioration of synthetic quality. The problem becomes serious during the voiced/unvoiced transition since the pitch estimation is prone to error at this region. The deficiency related to the pitch estimation can be alleviated by applying a median filter or error correction method to the pitch contour so that pitch halving, doubling as well as other deviating results can be avoided. To ameliorate the problem of a strict voicing decision, we propose a method to smooth the voicing transition as follows:

Consider a voiced segment, $s_v(i)$, that is next to an unvoiced segment. If the voiced segment is ahead of an unvoiced segment, then we can gradually change the voicing model by

$$s(i) = \frac{N_f + 1 - i}{N_f + 1} s_v(i) + \frac{i}{N_f + 1} s_{v(u)}(i), \qquad i = 1, 2, 3, \ldots, N_f \quad (4\text{–}44)$$

where $s(i)$ is the resultant speech signal, N_f is the frame length, and $s_{v(u)}(i)$ is an alternative version of $s_v(i)$ synthesized by using the innovation sequences. If the voiced segment is located behind an unvoiced segment, $s(i)$ becomes

$$s(i) = \frac{i}{N_f + 1} s_v(i) + \frac{N_f + 1 - i}{N_f + 1} s_{v(u)}(i), \qquad i = 1, 2, 3, \ldots, N_f \quad (4\text{–}45)$$

For simplicity, $s_{v(u)}(i)$ is derived at the analysis stage in our program. Nonetheless, it can

also be done at the synthesis stage where the CELP analysis-by-synthesis procedure is brought in to resynthesize $s_p(i)$ produced using the glottal impulse.

4.4 Subjective Quality Evaluation

Like many other waveform coders, the aim of the proposed source model is to extract important features that are not modeled by the LP filter. However, it is important to point out that although our synthetic speech waveform is very close to the original, we do not apply any closed-loop waveform-matching criterion nor a spectral weighting function while synthesizing voiced speech. It appears that the subjective measure on the basis of segmental SNR is not appropriate to indicate the quality of synthetic speech. For this reason, we conducted informal listening tests to assess the performance of the proposed source model as well as the LP speech synthesizer.

In addition to the training sentences, two other sentences have been tested, namely, "That zany van is azure" and "Should we chase those cow boys." The speech tokens included those uttered by speakers not in the training group. It was found that the quality of the synthetic speech was very close to that of the original speech. If the recorded speech was played back by loudspeakers in an A–B test, listeners found it difficult to discriminate the synthetic speech from its original counterpart. For speech tokens in which pitch contours were identical, the probability was approximately one-third that the synthetic speech were preferred over the original speech.

To acquire a more critical view of the excitation model, the listening tests were also carried out using a high-quality headphone (Sony, MDR–V6). It was revealed that our voiced excitation model tended to deliver more energy around the fundamental frequency, and the inverse filter could not fully counteract such a tendency. As a result, the synthetic speech was judged to be slightly bassy. However, when we increased the order of the inverse filter, the synthesized quality became crisper. Because such a crispy quality was not always preferred by listeners, we did not consider the increased order as an acceptable amelioration.

In contrast to the finite achievement of the inverse filter, both noise and source-tract interaction were more likely to be responsible for the improvement of synthetic quality. The addition of noise, in general, reduced the metallic attribute of synthetic speech. However, the use of a different amplitude modulation did not affect the speech quality largely. This is probably because the noise power for the modal voices is too low to result in any significant difference. The quality improvement due to the incorporation of source-tract interaction was noticeable in our experiments. We reason that as the combined result of raising formant resonances and attenuating the inter-formant components of the glottal impulse, which contributes the dispersiveness of the formant ripples and, in turn, reflects the fact that the residue is somewhat intelligible. From the view of spectral shaping, the resultant effect of formant ripples is considered the same as the amplitude spectrum modification introduced by Kang and Everett (1985) and the adaptive postfilter suggested by Chen and Gerso (1987). For this reason, a backward filtering operation that disperses the formant ripples ahead of the glottal closure is also recommended. In order words, the $W(z)$ that we used to modify the excitation pulse can be a zero-phase filter.

From the listening test, it was also found that modification with respect to g(3), g(4) and g(6) varied the pattern of vocal fold closure. According to our experience, a different closure pattern might lead to a change of the perceived quality. Although our empirical formula for constructing the excitation pulse suits a variety of voices, at present we do not have an appropriate theory to explain this result and to optimize the closure pattern.

Buzziness was reported in some synthetic speech of female speakers, especially for females with high fundamental frequencies of voicing. Using a visual comparison presented in Figure 4–9, we observe that the synthetic speech waveform for female voices, in general, has a direct bearing on the the synthetic excitation, which shows a more rapid uprising slope at the GCI and less noisy components during the glottal open phase. This leads us to suspect that the vocal fold closure pattern that is suitable for male voices may be too strong for some female voices. Likewise, the level of vocal noise for males may not be appropriate for

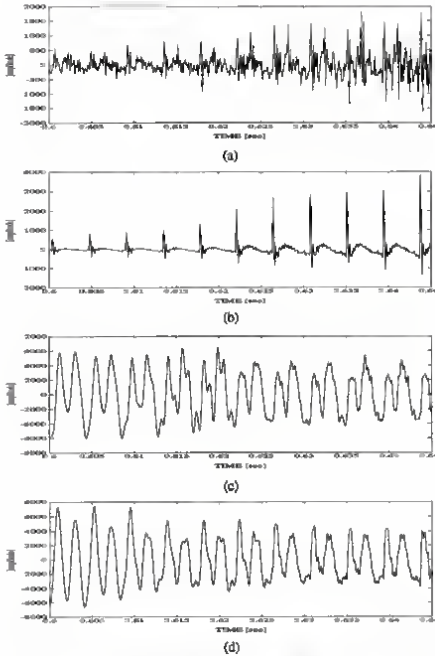


Figure 4–9 Comparison of the excitation and speech signals for a segment of voiced speech uttered by a female: (a) ideal excitation (residue). (b) synthetic excitation (glottal impulses), (c) original speech, (d) synthetic speech.

females. Essentially, an ideal glottal impulse should possess certain flexibilities in mixing the periodic pulses and vocal noise while preserving the necessary peakiness, the spectral tilt of the harmonic spectrum and the intensity of the fundamental component.

Roughness was also occasionally perceived as a degradation of female synthetic speech. Since the pitch irregularity has been considered to be an important correlate of roughness, the listening test results indicate the imperfection of our GCI identification algorithm, which relied on a sharp negative peak to have a proper initialization and consistent similarities of adjacent pitch periods to capture the real GCI's. In case this peak was smeared by nonstationary turbulence, the incorrect GCI's resulted in a domino effect at the latter stages that even our pitch smoothing procedure could not fully counteract.

Other perceivable distortions occurred in segments containing fricatives and nasal consonants. This implies that our excitation model can only partially replicate the spectral zeros (anti-formants). Because the observed phase characteristics of the nasals are not significantly different from that of the vowels, by inference, nasal sounds are not necessarily required in the codebook training. This inference has been further confirmed by testing the glottal codebook trained without nasals. No significant degradation was found for synthesized speech using such a codebook.

CHAPTER 5
CONCLUDING REMARKS

5.1 Summary

We confronted several problems in the first phase of this research. Attempts were made to verify the relationship between the residue signal and the glottal flow waveform. We concluded that the vocal characteristics could be removed from the integrated residue, which resembled the differentiated glottal flow. Also, within the source-filter theory, we proposed a comprehensive speech model that would encompass acoustic features previously used as quality attributes. The role of each model parameter was interpreted in the context of the acoustic measure. Thus, by making use of the LP analysis with the aid of EGG signals, we proposed methods for isolating and extracting the acoustic features. In particular, the perturbations of vocal source were decomposed into low-frequency drifts and wideband noise, where the latter was extracted by using a DFT method and later applied to derive the %jitter and %shimmer defined by Eqs. (2–15) and (2–16). The glottal spectral tilt was estimated using LP analysis of speech signals. While the removal of the spectral characteristics was performed by inverse filtering, the glottal phase was described by the waveshape of the integrated residue and a novel measure called abruptness index. The vocal noise was extracted from the integrated residue using a time domain approach and examined in three aspects, i.e., the relative power level, the amplitude modulation, and the noise spectrum. All the above mentioned feature extraction methods were demonstrated using sustained vowels /i/'s of three voice types (modal, vocal fry and breathy voices) as examples. The outcomes of these acoustic measures were carefully investigated and we have reached the following conclusions.

(1) As listed in Table 2–3, the distributions of %jitter and %shimmer for three voice types generally agreed with other researchers' results. More important, these results substantiated our assumptions that the perturbation noise exhibited a Gaussian distribution in which the standard deviation was sufficient to characterize the statistical property. If we consider the quality in a broader perspective, the gross pitch and intensity variations of speech signals not only transmit linguistic messages but also non-linguistic information such as intonation, emotional stress, and speaker idiosyncrasy. In order to synthesize natural speech, an accurate and faithful replication of these variations is necessary.

(2) The turbulent noise in breathy voices was perceptually distinctive and acoustically discriminable from that in the other two voice types. This underscores the need for a vocal noise model in the source model. The noise spectra for different phonations were fairly flat and therefore a white noise was suitable for modeling the vocal noise. On the other hand, although the amplitude modulations of the vocal noise generally resembles the magnitudes of the integrated residues, we proved that this result could result from the phase misalignment.

(3) The estimation of the glottal spectral tilt using LP analysis on the speech signal was tested with satisfactory results. In addition to visual inspection of the magnitude spectra of modeled filters, a simple comparison can also be carried out by testing the first coefficient of the underlying filter. The spectral tilts are moderate, relative flat, and steep for modal, vocal fry and breathy voices, respectively.

(4) The glottal phase characteristics did not show any significant relation across different voice types, suggesting no general rules for modeling the phase characteristics for different voice types. The abruptness index, in contrast, showed great potential for discriminating voice types, because the associated measures for each voice type are highly self-clustered and well separated from one another.

The above results provide a general idea of glottal variabilities. More extensive investigations are needed to establish the statistical significances between model parameters

and vocal quality The LP analysis appears to be capable of extracting the vocal source properties as well as the formant patterns. Thus, it is reasonable for us to argue that a high quality LP synthesizer is achievable if the acoustic features are accurately estimated and faithfully reproduced

In the second phase of this research, we were interested in the design of a high-quality natural-sounding LP synthesizer In Chapter 3, we presented an excitation model to simulate the voiced residue by the glottal impulses and the unvoiced residue by the innovation sequences. These two types of excitations were further formulated as two codebooks geared to an all-pole filter, of which the coefficients are estimated using the orthogonal covariance method. Schemes for speech analysis and synthesis were discussed in Chapter 4 Experiments with this new model and processing schemes demonstrated the competency of producing natural sounding speech In addition to source modeling, we believe efforts that lead to such encouraging results include the methods and algorithms performing the GCI identification, codeword searching, piece-wise LP interpolation, glottal pulse smoothing, spectral adjustment, source-tract interaction and gain determination These are either introduced for the first time in the literature or have had some modifications. Our achievements can be appreciated by appraising the quality of synthetic speech.

5.2 Possible Improvements

Though our LP synthesizer has been tested with fairly high success, there is still room for extension Several possible improvements are suggested as follows.

5.2.1 Extraction of Vocal Noise

Our noise extraction algorithm was impeded by the difficulty of phase misalignment. While the pitch delay is always restricted to integer multiples of the sampling (or resampling) interval, a possible method for overcoming this drawback is the use of a pitch predictor, for it not only provides the necessary interpolation but also maximizes the correlation between

the analyzed signals. In general, the number of filter taps need not be too many and the associated coefficients can be easily obtained by minimizing the mean squared error between the two signals. However, since the noise must be measured at the level of source excitation, more studies should be made concerning the effect of the sequential order of the pitch predictor and the inverse filter. Furthermore, as we already pointed out in Section 2.7, there are two types of noise presented in the residue signal, namely, the noise associated with the epoch variation and that with the airflow turbulence. One may consider how to decompose the residue signal into two such components, thus forcing the pitch predictor to examine the epoch or airflow variations separately. This will allow us to enlarge our view of the vocal noise.

5.2.2 GCI Identification

The improvement of performance and reliability of the GCI identification algorithm becomes an urgent requirement for high quality speech synthesis. In this research, we located the GCI's by first choosing the largest negative peak of the integrated residue in a frame as a reference mark and then searching for the other peaks by a maximum correlation approach. The resultant synthetic speech suffers probable distortion emerging from the inaccurate pitch identification. Thus, we have to rely on some correctional procedures to rectify some errant pitch transitions. It was reported that the GCI identification could achieve good performance if the maximum correlation approach was directly applied to the speech signal (Cheng and O'shaughnessy, 1989). Although our experiments with Cheng's approach did show some promising results, this approach has to be further refined before it can function automatically.

5.2.3 Excitation Source

Meanwhile, we are concerned with the excitation function for normal voices, leaving much latitude for the modification of the glottal codebook. For the utterance with elongated

glottal open phases the increased air turbulence will certainly perturb the accuracy of the inverse filter. Therefore, not only is the primary residue pulse less distinctive, but also there are other spurious components. It is obvious this type of residue cannot be properly described by a sharp glottal impulse. Also, based upon our perceptual impression, we believe that the strong excitation pulses should at least be partially responsible for the buzzy characteristics of female synthetic speech. One may think about ameliorating such defects by raising the vocal noise. According to our experience, adding noise did increase the breathiness but could not actually soften the voices. We therefore have to resort to other means. Several efforts in the past had been directed to designing excitation signals with low peak factors and flat magnitude spectra (Schroeder, 1970; Rabiner and Crochiere, 1975). Apparently, a logical follow-up of this research will be to consider designing a different codebook or developing processing schemes that control the peak factor and the vocal noise as well.

5.2.4 Ripple Effect

The ripple effect is known to be important for the improvement of speech quality, but it can only be simulated empirically in our experiments due to the lack of an efficient method for estimating the formant damping. What is the proper amount of the formant ripple that should be incorporated into the speech synthesizer in order to produce natural sounding speech? If we believe that the changes in formant damping only occur at the transitions from open-to-closed or closed-to-open glottis, then the utilization of techniques developed for fast formant tracking (Ting and Childers, 1990) and for the estimation of exponentially damped sinusoids (Parthasarathy and Tufts, 1987) may offer answers to this question. More studies are needed to decide how to control the damping factors and how these variations affect the quality. Even though the estimation procedures may be computationally prohibited from practical use, we would at least gain some qualitative description that might characterize different groups of speakers.

5.2.5 Sampling Resolution

Two approaches are considered useful to improve the sampling resolution. One is the use of a multiple-tap pitch predictor that we discussed in Section 5.2.1. The other is the fractional interpolation mentioned in Section 2.5.4. We believe the interpolated values of the signal will be able to represent the voiced speech more accurately and achieve an improvement of the synthetic quality for female speakers.

5.2.6 Spectral Estimation

In the proposed LP speech synthesizer, the spectra of speech signals are represented by an all-pole model. This model, however, is not suited for nasals and consonants, for the spectral envelopes of such sounds exhibit dips (zeros) besides peaks (poles). Though our source excitation partially compensates for the absence of spectral zeros by exploiting the adaptive nature of codeword searching, an Autoregressive Moving Average (ARMA) model seems to be even more attractive and straightforward for improving the quality of synthetic speech when spectral zeros are perceptually important (Atal and Schroeder, 1978, Akamine and Kiseki, 1989).

5.3 Applications

There are at least four areas where the techniques developed in this research are applicable.

5.3.1 Quality Measure

In this research we have postulated that a comprehensive speech production model should be constituted by a complete set of acoustic features. Thus, by utilizing the feature extraction techniques developed in Chapter 2, researchers would be able to determine a perceptually objective quality (or distortion) measure in a full context by verifying the relationships between the psychoacoustic attributes and the model parameters. Two possible

approaches that may achieve the described objective measure are: (1) statistical analysis based on large amounts of data, and (2) the analysis-by-synthesis procedure

In addition to assessing the speech quality (or severity), the resultant distortion measure can be used to study the speaker-characterizing acoustic features so that the speaker identification techniques can be advanced. Such a measure can also be used to study the way the human processes the acoustic signals so as to improve the speech recognition techniques. Furthermore, for any laryngeal pathology that is acoustically perceptible, the distortion measure can provide insight toward the classification of the pathologies

5.3.2 Speech Coding

We have not spent much effort dealing with the process of quantization and coding, but one may anticipate the superiority of the proposed model in speech coding Some advantages can be easily figured out by comparing the size of codebook for the various methods. The size of the glottal codebook required in our model is far less than that used by the DoD 4 8 Kb/s voice coder (Campbell et al , 1989) In addition, only one codeword is reserved to characterize the residue signal for each frame This will greatly reduce the demand of bit allocation and accelerate the processing speed

5.3.3 Voice Conversion

Because every codeword represents a different pattern of the glottal phase characteristics, the codeword searching can be considered as a process of monitoring and tracing the phase variation This implies that the glottal codebook can also be used to study the phase properties. As an example we have applied one single codeword to synthesize an entire sentence while keeping other parameters unchanged. The synthesized speech was intelligent but not as natural as the one synthesized by using the selected codeword sequence. Evidently, the poor quality is caused by the fixed glottal phase. This fact also indicates another possible application of this source model which is unfeasible by other waveform

coders. That is, our model can be used to convert the glottal phase characteristics. In the past, the LP synthesizers were only used to perform the prosodic and spectral modifications in the voice conversion systems (Childers et al., 1989b, Savic and Nam, 1991, Valbret et al., 1992). The acoustic realization using our model will render a complete version for voice conversion.

5.3.4 Text-to-Speech Synthesizer

The text-to-speech system is considered as the hierarchy of converting the text to a sequence of phonetic transcriptions before producing the acoustic output, which includes semantic, synthetic and lexical rules that monitor various intermediate transformation (Klatt, 1987). The final phonetic transcription is represented by the formant patterns (or LP coefficients), duration, pitch and intensity contours, all of which correspond the parameters of a speech production model. Thus, if the LP coefficient vectors for various phonetic segments were already registered in a codebook, we can easily apply our excitation codebooks to produce speech signals with desired quality. In fact, nowadays many LP codebooks are used for speech coding and have demonstrated high performance. It seems that the incorporation of this LP codebook into our current model is not only hypothetically feasible but also is a simple way to integrate the work of speech synthesis.

REFERENCES

Acheson, D J (1990) *Elementary Fluid Dynamics* (Oxford University Press, New York).

Ahn, C (1991) "A study of voice types and acoustic variability Analysis by synthesis," Ph.D Dissertation, University of Florida, Gainesville.

Akaike, H (1974) "A new look at the statistical model identification," IEEE Trans. Auto Control AC–19, 716–723

Akamine, M., and Kaseki K. (1989) "ARMA model based speech coding at 8kb/s," Proc IEEE Int. Conf. Acoust., Speech, Signal Process. 148–151

Allen, D. R , and Strong, W J (1985) "A model for the synthesis of natural sounding vowels," J Acoust. Soc Am 78(1), 58–69

Allen, J B (1977) "Short term spectral analysis and synthesis and modification by discrete Fourier transform," IEEE Trans. Acoust., Speech, Signal Process 25(3), 235–238

Allen, J. B., and Rabiner, L. R. (1977) "A unified theory of short-time spectrum analysis and synthesis," Proc IEEE, 65, 1558–1564

Almeida, L. B., and Silva, F. M (1984) "Variable-frequency synthesis An improved harmonic coding scheme," Proc IEEE Int. Conf Acoust., Speech, Signal Process. 27.5.1–27.5.4.

Almeida, L., and Tribolet, J M. (1982) "Harmonic coding A low bit-rate, good quality speech coding technique," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 1664–1667.

Ananthapadmanabha, T V., and Yegnanarayana, B (1979) "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoust., Speech, Signal Process 27(4), 309–319.

Askenfelt, A. G., and Hammarberg, B (1986) "Speech waveform perturbation analysis A perceptual-acoustical comparison of seven measures," J. Speech and Hear Res 29, 50–64

Atal, B S., and David, N (1979) "On synthesizing natural sounding speech by linear prediction," Proc IEEE Int. Conf Acoust., Speech, Signal Process 44–47.

Atal, B. S., and Hanauer, S. L. (1971) "Speech analysis and synthesis by linear prediction of the speech wave," J Acoust. Soc Am 50(2), 637–655

Atal, B. S , and Remde, J R. (1982) "A new model of LPC excitation for producing natural sounding speech at low bit rates," Proc IEEE Int. Conf Acoust. Speech Signal Process. 614–617.

Atal, B. S., and Schroeder, M. R. (1978). "Linear prediction analysis of speech based on a pole-zero representation," J. Acoust. Soc. Am. 64(5), 1310–1318.

Atal, B. S., and Schroeder, M. R. (1979). "Predictive coding of speech signals and subjective error criteria," IEEE Trans. Acoust., Speech, Signal Process. 27(3), 247–254.

Berg, J. W. van den (1958). "Myoelastic-aerodynamic theory of voice production," J. Speech and Hear. Res., 1, 227–244.

Berg, J. W. van den, Zantema, J. T., and Doornenbal, P. Jr. (1957). "On the air resistance and the Bernoulli effect of the human larynx," J. Acoust. Soc. Am. 29, 626–631.

Bergstrom, A., and Hedelin, P. (1989). "Code-book driven glottal pulse analysis," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 53–56.

Berouti, M. G. (1976). "Estimation of glottal volume velocity by the linear prediction inverse-filter," Ph.D. Dissertation, University of Florida, Gainesville.

Bladon, R. A. W., and Lindblom, B. (1981). "Modeling the judgement of vowel quality differences," J. Acoust. Soc. Am. 69(5), 1414–1422.

Buzo, A., Gray, A. H. Jr., Gray, R. M., and Markel, J. D. (1980). "Speech coding based upon vector quantization," IEEE Trans. Acoust., Speech, Signal Process. 28, 562–274.

Campbell, J. P., and Thomas, E. T. (1986). "Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10e algorithm," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 473–476.

Campbell, J. P., Welch, V. C., and Tremain, T. E. (1989). "An expandable error-protected 4800 bps CELP coder," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 735–738.

Caspers, B., and Atal, B. S. (1987). "Role of multi pulse excitation in synthesis of natural-sounding voiced speech," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2388–2391.

Chandra, S., and Lin, W. C. (1974). "Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis," IEEE Trans. Acoust., Speech, Signal Process. 22(6), 403–415.

Chen, J. H., and Gersho, A. (1987). "Real-time vector APC speech coding at 4800 bps with adaptive postfiltering," Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. 2185–2188.

Cheng, Y. M., and O'Shaughnessy, D. (1989). "Automatic and reliable estimation of glottal closure instant and period," IEEE Trans. Acoust., Speech, Signal Processing, 37(12), 1805–1815.

Childers, D. G., and Bae, K. S. (1992). "Detection of laryngeal function using speech and electroglottographic data," IEEE Trans. on Biomedical Eng. 39(1), 19–25.

Childers, D. G., Hahn, M., and Larar, J. N. (1989a). "Silent and voiced/unvoiced/mixed excitation (four-way) classification of speech," IEEE Trans. Acoust., Speech, Signal Process. 37(11), 1771–1774.

Childers, D G., and Larar, J N. (1984). "Electroglottography for laryngeal function assessment and speech analysis," IEEE Trans. on Biomedical Eng 31(12), 807–817

Childers, D G , and Lee, C K. (1991) "Vocal quality factors. Analysis, synthesis, and perception," J Acoust. Soc Am. 90, 2394–2410

Childers, D G , and Wu, K. (1990). "Quality of speech produced by analysis-synthesis," Speech Commun 9, 97–117

Childers, D G Wu K., Hicks, D M , and Yegnanarayana, B (1989b) "Voice conversion," Speech Commun. 8, 147–158

Childers, D G , Yea, J J., and Krishnamurthy, A. (1981) "Spectral analysis: AR, MA, ARMA," First IEEE Acoust., Speech, Signal Process. Workshop on Spectral Estimation, 2.2.1–2.2.10

Childers, D. G , Yegnanarayana, B., and Wu, K. (1985) "Voice conversion: Factors responsible for quality," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 748–751

Colton, R H (1973) "Vocal intensity in the modal and falsetto registers," Folia Phoniatrica 25, 62–70

Dankberg, M. D , and Wong, D Y (1979) "Development of a 4.8–9 6 kbps RELP vocoder," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 554–557

Deller, J R. (1982) "Evaluation of laryngeal dysfunction based on features of an accurate estimate of the glottal waveform," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 759–762.

Deller, J R. and Anderson, D J (1980) "Automatic classification of laryngeal dysfunction using the roots of the digital inverse filter," IEEE Trans. on Biomedical Eng. 27, 714–721

Dudley, H (1939). "The vocoder" Bell Labs Rec. 18, 122–126.

Eskenazi L., Childers, D G. and Hicks, D M (1990) "Acoustic correlates of vocal quality" J Speech and Hear Res. 33, 298–306

Fant, G. (1959) "The acoustics of speech," Proc. 3rd Int. Cong. on Acoust. 188–201

Fant, G. (1960) *Acoustic Theory of Speech Production* (Mouton, Paris)

Fant, G , and Ananthapadmanabha, T V (1982) "Truncation and superposition" Speech Trans Lab –Q Prog. Status Rep. (Royal Institute of Technology, Stockholm, Sweden) 2–3, 1–17

Fant, G., Liljencrants, J and Lin, Q (1985) "A four-parameter model of glottal flow," Speech Trans. Lab –Q Prog. Status Rep 4, 1–13

Fant, G., and Lin, Q (1988) "Frequency domain interpretation and derivation of glottal flow parameters," Speech Trans. Lab –Q Prog. Status Rep 2–3, 1–21

Flanagan, J. L. (1972a) *Speech Analysis, Synthesis and Perception* (Springer Verlag, New York), 2nd ed.

Flanagan, J. L. (1972b) "Voices of man and machines," J. Acoust. Soc. Am. 51(5), 1375–1387

Flanagan, J. L., and Golden, R. M. (1966) "Phase vocoder," Bell Syst. Tech. J., 45, 1493–1509

Fujisaki, L, and Ljungqvist, M. (1986) "Proposal and evaluation of models for the glottal source waveform," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 1605–1608

Fukazawa, T., El-Assuoofy, A., and Honjo, I. (1988) "A new index for evaluation of the turbulent noise in pathological voice," J. Acoust. Soc. Am. 83, 1189–1193

Furui, S. (1985). *Digital Speech Processing, Synthesis, and Recognition* (Marcel Dekker, New York)

Galand, C. R., Menez, J. E., and Rosso, M. M. (1992). "Adaptive code excited predictive coding," IEEE Trans. Signal Process. 40(6), 1317–1326

Gobl, C. (1988). "Voice source dynamics in connected speech," Speech Trans. Lab. –Q. Prog. Status Rep. 2–3, 123–159

Gobl, C. (1989) "A preliminary study of acoustic voice quality," Speech Trans. Lab. –Q. Prog. Status Rep. 4, 9–22.

Gray, R. M. (1984) "Vector quantization," IEEE ASSP Magazine (April), 4–29

Griffin, D. W., and Lim, J. S. (1988) "Multiband excitation vocoder," IEEE Trans. Acoust., Speech, Signal Process. 36(8), 1223–1235

Guérin, B., Mrayati, M., and Carré, R. (1976) "A voice source taking account of coupling with the supraglottal cavities," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 47–50

Haagen, J., Nielsen, H., and Hansen S. D. (1992). "Improvements in 2.4kps high-quality speech coding," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. II-145–II-148

Hedelin, P. (1981) "A tone-oriented voice-excited vocoder," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 205–208.

Hedelin, P. (1986). "High quality glottal LPC vocoding," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 465–468

Hedelin, P. (1988) "Phase compensation in all-pole speech analysis," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 339–342.

Hermansky H., Hanson, B. A., and Wakita, H. (1985) "Perceptually based linear predictive analysis speech," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 509–512.

Hiki, S., Imaizumi, S., Hirano, M., Matsushita, H., and Kakita, Y. (1976). "Acoustical analysis for voice disorders," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 613–616.

Hillman, R. E., and Weinberg, B. (1981) "Estimation of volume velocity waveform properties. A review and study of some methodological assumptions," in *Speech and Language: Advances in Basic Research and Practice*, edited by N. Lass (Academic P., New York), pp. 411–473

Hiraoka, M , Kitazoe, Y , Ueta, H., Tanaka, S , and Tanabe, M (1984) "Harmonic-intensity analysis of normal and hoarse voices," J Acoust. Soc Am. 76, 1648–1651

Hollien, H. (1974) "On vocal register" J Phon 2, 125–144

Holmes, J N (1973) "The influence nf glottal waveform on the naturalness of speech from a parallel formant synthesizer," IEEE Trans Audio Electroacout. AU–21(3), 298–305

Holmes, J N (1983) "Formant synthesizers, cascade or parallel," Speech Commun. 2 251–274

Javkin, H. R., Antonanzas-Barroso, N., and Maddieson, I. (1987) "Digital inverse filtering for linguistic research," J Speech and Hear Res 30, 122–129

Kahn, M , and Garst, P (1983). "The effects of five voice characteristics on LPC quality," Proc IEEE Int. Conf Acoust., Speech, Signal Process. 531–543

Kang, G S., and Everett, S (1985) "Improvement of the excitation source in the narrow-band linear prediction vocoder," IEEE Trans. Acoust., Speech, Signal Process. 33(2), 377–386

Karlsson, I (1988) "Glottal waveform parameters for different speaker types," Speech Trans Lab –Q Prog. Status Rep 2–3, 61–67

Kasuya, H., Ogawa, S , and Kikuchi, Y (1986a) "An acoustic analysis of pathologic voice and its application to the evaluation of laryngeal pathology," Speech Commun 5, 171–181

Kasuya, H., Ogawa, S., Mashima, K., and Ebihra, S (1986b) "Normalized noise energy as an acoustic measure to evaluate pathologic voice," J Acoust. Soc. Am. 80(5), 1329–1334

Kitajima, K. (1981) "Quantitative evaluation of the noise level in the pathological voice," Folia Phoniatrica, 33, 115–124.

Klatt, D H. (1980) ' Software for a cascade/parallel formant synthesizer," J Acoust. Soc Am 67(3), 971–995

Klatt, D H. (1987) "Review of text-to-speech conversion for English," J Acoust. Soc Am 82(3), 737–793

Klatt, D. H., and Klatt, L. C. (1990) "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J Acoust. Soc Am 87(2), 820–857

Kleijn, W B , Krasinski, D J , and Ketchum, R. H. (1990) "Fast methods for the CELP speech coding algorithm," IEEE Trans. Acoust., Speech Signal Process. 38(8) 1330–1342.

Koike, Y (1969) ' Vowel amplitude modulation in patients with laryngeal diseases, J Acoust. Soc Am 45(4). 839–844

Koike, Y., and Markel, J (1975) "Application of inverse filtering for detecting laryngeal pathology," Annals of Otology, Rhinology and Laryngology 84(1), 117–124

Konstantinides, K. (1991) "Threshold bounds in SVD and a new iterative algorithm for order selection in AR models," IEEE Trans. Signal Process. 39(5), 1218–1221.

Konstantinides, K., and Yao, K. (1988). "Statistical analysis of effective singular values in matrix rank determination." IEEE Trans. Acoust., Speech, Signal Process. 36(5), 757–763.

Krishnamorthy, A. K., and Childers, G. C. (1986) "Two-channel speech analysis," IEEE Trans Acoust. Speech Signal Process. 34(4). 730–743.

Kroon, P., and Atal, B. S. (1990) "Pitch predictors with high temporal resolution," Proc IEEE Int. Conf. Acoust., Speech, Signal Process. 661–664.

Kuwabara, H. (1984). "A pitch-synchronous analysis/synthesizer system to independently modify formant frequencies and bandwidths for voiced speech," Speech Commun. 3, 211–220.

Kwon, S. Y., and Goldberg, A. J. (1984) "An enhanced LPC vocoder with no voiced/unvoiced switch," IEEE Trans. Acoust., Speech Signal Process. 32(4), 851–858.

Lalwani, A. L., and Childers, D. G. (1991) "Modeling vocal disorders via formant synthesis," Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 505–508.

Laver, J., and Hanson, R. (1981) "Describing the normal voice," in *Evaluation of Speech in Psychiatry* edited by J. Darby (Grune and Stratton, New York), pp. 51–78.

Lieberman, P. (1961) "Perturbation in vocal pitch," J. Acoust. Soc. Am. 33(5). 597–603.

Lieberman, P., and Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge U. P., New York).

Linde, Y., Buzo, A., and Gray, R. M. (1981) "An algorithm for vector quantizer design," IEEE Trans. Commun. COM-28(1), 84–95.

Ma, C. K., and Chan, C. K. (1991) "Maximum decent method for image vector quantisation," Electronics Letters. 27(19), 1772–1773.

Makhoul, J. (1975) "Linear prediction: A tutorial review," Proc. IEEE, 63, 561–580.

Makhoul, J., Viswanathan, R., Schwartz, R., and Huggins, A. W. F. (1978) "A mixed-source model for speech compression and synthesis," J. Acoust. Soc. Am. 64(6), 1577–1581.

Markel, J. D., and Gray, A. H. (1976) *Linear Prediction of Speech* (Springer-Verlag, New York).

Matausek, M., and Batalov, V. (1980) "A new approach to the determination of the glottal waveform," IEEE Trans. Acoust., Speech, Signal Process. 28(6), 616–622.

McAulay, R. J., and Quatieri, T. F. (1984). "Magnitude-only reconstruction using a sinusoidal speech model." Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 27.6.1–27.6.4.

McCree, A. V., and Barnwell, T. P. (1991) "A new mixed excitation LPC vocoder," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 593–596.

Milenkovic, P (1986) "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. Acoust., Speech, Signal Process. 34(1), 28–42

Milenkovic, P. H. (1987). "Least mean square measures of voice perturbation," Journal of Speech and Hear. Res. 30, 529–538

Monsen, R., and Engebretson, M. (1977). "Study of variations in the male and female glottal wave," J. Acoust. Soc. Am. 62(4), 981–993

Moore, G. P (1976). "Observation on laryngeal disease, laryngeal behavior, and voice," Annals of Otology, Rhinology and Laryngology 85, 553–567

Muta, H., Muraoka, T., Wagatsuma, K., Fukuda, H., Takayama, E., Fujioka, T., and Kanou, S. (1987) "Analysis of hoarse voices using the LPC Method," in *Laryngeal Function in Phonation and Respiration*, edited by T. Baer, C. Sasaki, and K. Harris (College-Hill Press, Boston, MA), pp. 463–474

Nung, T., and Whiting, S. (1990) "Power spectrum estimation via orthogonal transformation," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2523–2526

Nyeck, A., and Tosser-Roussey, A. (1992) "Maximum entropy initialisation technique for image coding vector quantiser design," Electronics Letters. 28(3), 273–274

Oppenheim, A. V. (1969) "A speech analysis-synthesis system based on homomorphic filtering," J. Acoust. Soc. Am. 45(2), 458–465

Oppenheim, A. V., and Willsky, A. S. (1983) *Signals and Systems* (Prentice-Hall, Englewood Cliffs, NJ)

Pao, Y.-H. (1989) Adaptive pattern recognition and neural networks (Addison-Wesley, New York)

Parthasarathy, S., and Coker, C. H. (1992) "On automatic estimation of articulatory parameters in a text-to-speech system," Computer Speech and Language 6, 37–75

Parthasarathy, S., and Tufts, D. W. (1987). "Excitation-synchronous modeling of voiced speech," IEEE Trans. Acoust., Speech, Signal Process. 35(9), 1241–1249

Pinto, N. B., Childers, D. G., and Lalwani, A. L. (1989). "Formant speech synthesis: Improving production quality," IEEE Trans. Acoust., Speech, Signal Process. 37(12), 1987–1887

Pinto, N. B., and Titze, I. R. (1990) "Unification of perturbation measures in speech signals," J. Acoust. Soc. Am. 87(3), 1278–1289

Pisoni, D. B., and Hunnicutt, S. (1980) "Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 572–575

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., and Schwab, E. C. (1983) "Perceptual evaluation of synthetic speech: Some considerations of the user/system interface," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 535–538

Prado, P. P. L., Shiva, E. H., and Childers, D. G. (1992). "Optimization of acoustic-to-articulatory mapping," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. II-33–11–36.

Price, P. J. (1989). "Male and female voice source characteristics: Inverse filtering results," Speech Commun. 8, 262–277.

Prosek, A. R., Montgomery, B. E., and Hawkins, D. B. (1987). "An evaluation of residue features as correlates of voice disorders," J. Commun. Disorders 20, 105–117.

Rabiner, L. R., and Crochiere, R. E. (1975). "On the design of all-pass signals with peak amplitude constraints," Bell Syst. Tech. J., 55(4), 395–407.

Rabiner, L. R., and Schafer, R. W. (1978). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).

Robinson, D., and Dadson, R. (1956). "A redetermination of the equal-loudness relations for pure tones," Br. J. Appl. Physics 7, 166–181.

Rose, R. C., and Barnwell, T. P. III. (1990). "Design and performance of an analysis-by-synthesis class of predictive speech coders," IEEE Trans. Acoust., Speech, Signal Process. 38(9), 1489–1503.

Rosenberg, A. E. (1971). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. 49(2). 583–590.

Rothenberg, M. (1981). "Acoustic interaction between the glottal source and the vocal tract," in *Vocal Fold Physiology*, edited by K. N. Stevens, and M. Hirano (Univ. of Tokyo Press), pp. 305–323.

Sambur, A. E., Rosenberg, L. R., Rabiner, L. R., and McGonegal, C. A. (1978). "On reducing the buzz in LPC synthesis," J. Acoust. Soc. Am. 63(3), 918–924.

Savic, M., and Nam, J.-H. (1991). "Voice personality transformation," Digital Signal Processing 1, 107–110.

Schafer, R. W., and Rabiner, L. R. (1973). "A digital signal processing approach to interpolation," Proceedings of the IEEE, 61, 692–702.

Schoentgen, J. (1982). "Quantitative evaluation of the discrimination performance of acoustic features in detecting laryngeal pathology," Speech Commun. 1, 269–282.

Schoentgen, J. (1989). "Jitter in sustained vowels and isolated sentences produced by dysphonic speakers," Speech Commun. 8, 61–79.

Schroeder, M. R. (1970). "Synthesis of low-peak-factor signals and binary sequences with low autocorrelation," IEEE Trans. Inform. Theory, IT-16, 85–89.

Schroeder, M. R., and Atal, B. S. (1985). "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 937–940.

Schultheib, M. and Lacroix, A. (1989). "On the performance of CELP algorithms for low rate speech coding," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 152–155.

Schumacher, R. T., and Chafe, C. D. (1990). "Characterization of aperiodicity in nearly periodic signals," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, 1161–1164.

Schwarz, G (1978). "Esumatioo of the dimeasion of the model ' Ann Stat. 6, 461–464

Singhal, S., and Atal, B S. (1989) "Amplitude optimization and pitch prediction in multipulse coders," IEEE Trans. Acoust., Speech, Signal Process. 37(3), 317–327

Smith, A. M , and Childers, D G (1983). "Laryngeal evaluation using features from speech and the electroglottograph," IEEE Trans. on Biomedical Eng. 30(11), 755–759

Sorensen, D , and Horii, Y. (1984) "Directional perturbation factors for jitter and shimmer," J Commuo Disorders 17, 143–151

Sreenivas, T V. (1988). "Modelling LPC-residue by components for good quality speech coding," Proc IEEE Int. Conf Acoust. Speech Signal Process. 171–174

Sukkar, R. A., LoCicero, J L., and Picone, J. W. (1989) "Decomposition of the LPC excitation using the Zinc basis functions," IEEE Trans Acoust., Speech, Signal Process 37(9), 1329–1341

Tenpaku, S., and Hirahara, T (1990) "A glottal waveform model for high-quality speech synthesis," J. Acoust. Soc. Am. 88, S152.

Ting, Y. T., and Childers, D. G. (1990) "Speech analysis using the weighted recursive least squares algorithm with a variable forgetting factor," Proc IEEE Int. Conf. Acoust., Speech, Signal Process. 389–392.

Titze, I. R., Horii, Y., and Scherer R. C. (1987) "Some technical considerations in voice perturbation measurements," J Speech and Hear Res 30 252–260

Tohkura, Y., Itakura, F., and Hashimoto, S (1978) "Spectral smoothing technique in PARCOR speech analysis-synthesis," IEEE Trans. Acoust., Speech, Signal Process. 26(6), 587–596

Tou, J T (1979) "DYNOC — A dynamic optimal cluster-seeking technique," Int. J Comput. Inf Sci. 8(6), 541–547

Tou, J. T., and Gonzalez, R. C. (1974) *Pattern Recognition Principles* (Addison-Wesley, New York)

Trancoso, I. M., and Atal, B S. (1990) 'Efficient search procedure for selecting the optimum innovation in stochastic coders," IEEE Trans. Acoust., Speech, Signal Process. 38(3), 385–396.

Trancoso, I. M , Marques, J S , and Ribeiro, C. M. (1990) "CELP and sinusoidal coders. Two solutions for speech coding at 4.8–9 6 kbps," Speech Commun. 9, 389–400.

Un C. K., and Magill, D T (1975) "The residual-excited linear prediction vocoder with transmissioo rate below 9 6 kbits/s," IEEE Trans. Commun. COM–23(12), 1466–1474.

Valbret, H., Moulines E., and Tubach, J P (1991). "Voice transformation using PSOLA technique," Speech Commun 11, 175–187.

Verhelst, W., and Nilens, P (1986). "A modified-superposition speech synthesizer and its application," Proc IEEE Int. Conf. Acoust., Speech, Signal Process. 2007–2010.

Viswanathan, R., and Makhoul, J. (1975) "Quantization properties of transmission parameters in linear predictive systems," IEEE Trans. Acoust., Speech, Signal Process. 23(3), 309–321

Wang, S., Sekey, A., and Gersho, A. (1991). "Auditory distortion measure for speech coding," Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. 493–496

Wax, M. (1988) "Order selection for AR models by predictive least squares," IEEE Trans. Acoust., Speech, Signal Process. 36(4). 581–588

Wendahl, R. W. (1963). "Laryngeal analog synthesis of harsh voice quality," Folia Phoniatrica 15, 241–250

Wolfe, V. I., and Steinfatt, T. M. (1987). "Prediction of vocal severity within and across voice types," J. Speech and Hear. Res. 30, 230–240.

Wong, C.-H. (1991) "The incorporation of glottal source vocal tract interaction effects to improve the naturalness of synthetic speech," Ph.D. Dissertation, University of Florida, Gainesville.

Wong, D. Y. (1980) "On understanding the quality problems," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 725–728.

Wong, D. Y., and Markel, J. D. (1978). "An excitation function for LPC synthesis which retains the human glottal phase characteristics," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 171–174

Wong, D. Y. Markel, J. D., and Gray, A. H. Jr. (1979). "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Process. 27(4), 350–355.

Yea, J. J., Krishnamurthy, A. K. Naik J. K., Moore, G. P., and Childers, D. G. (1983) "Glottal sensing for speech analysis and synthesis," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 1332–1335

Yumoto, E., Gould, W., and Baer, T. (1982) "Harmonics-to-noise ratio as an index of the degree of hoarseness," J. Acoust. Soc. Am. 71, 1544–1550

Yumoto, E., Sasaki, Y., Okamura, H. (1984) "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness," J. Speech and Hear. Res. 27, 2–6.

Zhang, X., and Chang, X. (1992). "A new excitation model for LPC vocoder at 2.4 Kb/s," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. I-65–I-68.

Zinser, R. L., and Koch, S. R. (1989) "4800 and 7200 bit/sec hybrid codebook multipulse coding," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 747–750.

BIOGRAPHICAL SKETCH

Hwai-Tsu Hu was born in Gangsan, a town in south Taiwan, Republic of China, on January 15, 1964. He graduated from National Cheng Kung University, Tainan, Taiwan, in June, 1985 with a Bachelor of Science degree in electrical engineering. After complying with the compulsory military service, he entered the Department of Electrical Engineering, University of Florida, where he received his Master of Science degree in May, 1990. Since then, he has been a graduate research assistant under the supervision of Dr. D. G. Childers at the Mind/Machine Interaction Research Center, where his primary interest is digital signal processing with application to speech analysis and synthesis. After completing the requirements for the Ph.D. degree, he intends to return to his country and anticipates getting involved in research areas such as VLSI signal processing, digital image and speech processing, as well as electronic telecommunications.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald G. Childers

Donald G Childers, Chairman
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Leon W Couch, II
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Fred J Taylor
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jose C Principe
Associate Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Mark C. K. Yang
Professor of Statistics

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May, 1993

Winfred M. Phillips
Dean, College of Engineering

Madelyn M. Lockhart
Dean, Graduate School